SLTC 2012

The Fourth Swedish Language Technology Conference

Lund, October 24-26, 2012

Proceedings of the Conference

The organizers gratefully acknowledge the support from the Swedish Graduate School in Language Technology (GSLT), as well as our sponsors:

Platinum	FINDWISE SEARCH DRIVEN SOLUTIONS
Gold	Oribi
Gold	LUNDS KOMMUN
Silver	SONY

Cover photo: Pierre Mens
Layout: Marcus Uneson

Preface

After Göteborg (2006), Stockholm (2008), and Linköping (2010), Lund University is proud to host the 4th Swedish Language Technology Conference (SLTC) in the beautiful settings of the Centre for Languages and Literature (Språk- och litteraturcentrum, SOL) and the Lund Institute of Technology (Lunds tekniska högskola, LTH).

SLTC is a unique and friendly event that gathers most of the researchers active in the field in Sweden. This year, we received 53 submissions. We accepted 43 papers and 42 appear in the final proceedings. By reading these proceedings and attending the communications, we can be sure that SLTC is truly reflecting the vibrant activity in language technology in Sweden. We are also happy to welcome a number of international contributions.

We are extremely pleased to open our two conference days with keynote lectures from two outstanding researchers in language technology: Martha Palmer from the University of Colorado and Fred Roberts from Artificial Solutions.

As for the past SLTC conferences, we are delighted to accommodate more specialized symposia in three collocated workshops:

- Language, action, and perception;
- NLP for computer assisted language learning; and
- Exploratory query-log analysis.

SLTC is also collocated with the Nordic seminar on speech recognition organized by ASTIN (The workgroup for Language Technology in the Nordic Countries).

SLTC 2012 would not have been possible without the support of the Graduate School of Language Technology (GSLT), SLTC's parent organization. We also gratefully acknowledge our sponsors – Vetenskapsrådet, Findwise, Oribi, Lunds kommun, and Sony Mobile – and the enthusiastic cooperation of the main venue, the Centre for Languages and Literature. Their generous support has permitted to maintain very modest participation costs. Finally, we would like to thank the program committee that did an incredible job in reviewing all the papers on a very tight schedule.

We hope the participants will enjoy SLTC 2012, share and find new ideas, and that we will together make this conference a successful event.

Pierre Nugues,

on behalf of the organizing committee: Anders Ardö (LTH), Peter Exner (LTH), Jonas Granfeldt (SOL), Marcus Uneson (SOL), Caroline Willners (SOL and Oribi), Jonas Wisbrant (LTH).

Program Committee

Lars Ahrenberg, Linköping
Anders Ardö, Lund
Gerlof Bouma, Göteborg
Peter Exner, Lund
Karin Friberg Heppin, Göteborg
Natalia Grabar, Lille
Jonas Granfeldt, Lund
Thierry Hamon, Paris
Anna Hjalmarsson, Stockholm
Merle Horne, Lund
Richard Johansson, Göteborg
Sabine Kirchmeier-Andersen, Copenhagen
Pierre Nugues (chair), Lund
Gabriel Skantze, Stockholm
Sara Stymne, Linköping
Jörg Tiedemann, Uppsala
Marcus Uneson, Lund

Invited speaker: Martha Palmer, University of Colorado

Beyond Shallow Semantics

Shallow semantic analyzers, such as semantic role labelers and sense taggers, are increasing in accuracy and becoming commonplace. However, they only provide limited and local representations of words and individual predicate-argument structures.

This talk will address some of the current opportunities and challenges in producing deeper, richer representations of coherent eventualities. Available resources, such as VerbNet, that can assist in this process will also be discussed, as well as some of their limitations.

Speaker's Bio

Martha Palmer is a Full Professor at the University of Colorado with joint appointments in Linguistics and Computer Science and is an Institute of Cognitive Science Faculty Fellow. She recently won a Boulder Faculty Assembly 2010 Research Award.

Prof. Palmer's research has been focused on trying to capture elements of the meanings of words that can comprise automatic representations of complex sentences and documents. Supervised machine learning techniques rely on vast amounts of annotated training data so she and her students are engaged in providing data with word sense tags and semantic role labels for English, Chinese, Arabic, Hindi, and Urdu, funded by DARPA and NSF. They also train automatic sense taggers and semantic role labelers, and extract bilingual lexicons from parallel corpora. A more recent focus is the application of these methods to biomedical journal articles and clinical notes, funded by NIH.

Prof. Palmer is a co-editor for the Journal of Natural Language Engineering and for LiLT, Linguistic Issues in Language Technology, and on the CLJ Editorial Board. She is a past President of the Association for Computational Linguistics, past Chair of SIGLEX and SIGHAN, and was the Director of the 2011 Linguistics Institute held in Boulder, Colorado.

Invited speaker: Fred Roberts, Artificial Solutions

Teneo, a Visual Approach to Building Virtual Assistants

Since 1999, Artificial Solutions has built several hundred virtual assistants, and in the process, developed the science of dialogue construction to a fine art. Recently we have reinvented our platform to accommodate a more visual approach in building knowledge, making it easier to implement a wide range of dialogue requirements on the one hand, while making the knowledge more accessible to our customers, on the other hand.

In a demonstration of this technology, we will show some of the capabilities we build into our systems and offer a behind-the-scenes look at the Teneo platform to show how it supports multiple domains, dialogue management, disambiguation, multitasking, task resumption and other related concepts.

Speaker's Bio

Fred Roberts is an award-winning R&D Engineer at Artificial Solutions Germany working with NLI technology since 2000. He is creator and designer of Elbot, first place winner of the Chatterbox Challenge 2003 and the 2008 Loebner Competition. He has published several short stories and articles, and received the □Google blog of note□ citation for a personal Weblog. He has summa cum laude degrees in Computer Science and Psychology, studied at Northern Kentucky University, USA and Bielefeld University, Germany and is a former employee of IBM and Nixdorf. In 2010 he was named an NKU Outstanding Alumnus, College of Informatics.

Table of Contents

Processing spelling variation in historical text

Yvonne Adesam and Malin Ahlberg and Gerlof Bouma

Språkbanken
Institutionen för svenska språken
Göteborgs universitet
firstname.lastname@gu.se

1. Overview

Språkbanken, the Swedish language bank,[1] has a large collection of modern texts with close to a billion words, which have been automatically annotated and are searchable through an online corpus query interface (Borin et al., 2012). We are currently working on increasing the diachronic coverage of the corpora (Borin et al., 2010), by including Swedish texts from the 19th century back to the 13th century. Ultimately, our goal is to develop tools for all types of text, at various levels of annotation, such as part-of-speech, morphosyntactic information, and dependency parses.

Our primary source material for Old Swedish (ca 1225-1526) comes from Fornsvenska textbanken,[2] a collection of around 160 digitized texts, comprising 3 million words in total, mainly from the 13th to the 16th century. The data set consists of novels, poems, laws, and religious texts, ranging from 200 to 200,000 words in length. A number of problems arise when handling this older corpus material. This already starts at the very beginning of the pipeline, with for instance sentence splitting (boundary markers such as punctuation and capitalization are commonly missing) and defining types. The latter is caused by a lack of standardized orthography. To give an example, the type corresponding to 'letter.PL.INDEF' occurs in our corpus as *bokstaffua*, *bokstaffwa*, *bokstafwa*, *bokstaua*, *bokstawa*, *bogstaffua*. Not recognizing these as belonging to one type will increase data sparseness. Likewise, a tool that is not robust against different spellings, for instance, the computational Old Swedish morphology of Borin and Forsberg (2008), may be restricted in its direct applicability to new historical material.

This paper presents our ongoing work in handling Old Swedish spelling variation as a first part of our greater aim to provide processing tools for historical material.

2. Spelling variation

A common approach to spelling variation is to normalize all spelling to modern orthography (Jurish, 2010; Bollmann et al., 2011; Pettersson et al., 2012). For instance, the examples above could be normalized to modern orthography *bokstava* (although this form no longer exists as a plural nominal). In contrast, we focus on matching words in the texts to the lexical entries in three Old Swedish dictionaries: Söderwall and Söderwall's supplement (1884/1953) with 44 000 entries, and Schlyter (1887) with 10 000 entries, focusing on law texts. Approximate string matching is done through minimum edit distance matching. The edit distances are based upon weighted substitution rules that may apply to substrings of unequal length.

o	u	0.202	a#	#	0.338	ki	k	0.406
æ	a	0.247	au	ö	0.442	tt	t	0.274
y	i	0.255	r#	#	0.253	ll	l	0.363
ö	o	0.288	er#	#	0.339	nn	n	0.386
j	i	0.299	c	k	0.307	þ	d	0.247
ö	y	0.306	ft	pt	0.314	aþ	an	0.373
e	i	0.311	gh	g	0.344	þe	n	0.378
å	a	0.313	f	p	0.361	dh	þ	0.386

Table 1: Some example substitution rules with cost.

We automatically extracted approximately 6 000 weighted substitution rules from alternative entries in the Schlyter lexicon. See Adesam et al. (2012) for more details. Examples are shown in Table 1. The small pilot evaluation presented in that paper suggests that the rules increase the proportion of tokens matched from about 25% (exact matching) to almost 100%. At the same time, precision of the first lexicon match went down from 80% to 65%. Encouragingly, a correct match *is* found within the top 3 candidates for more than 80% of the cases. The high coverage of the substitution rules demonstrates that they also capture variation due to inflection, albeit as a side effect.

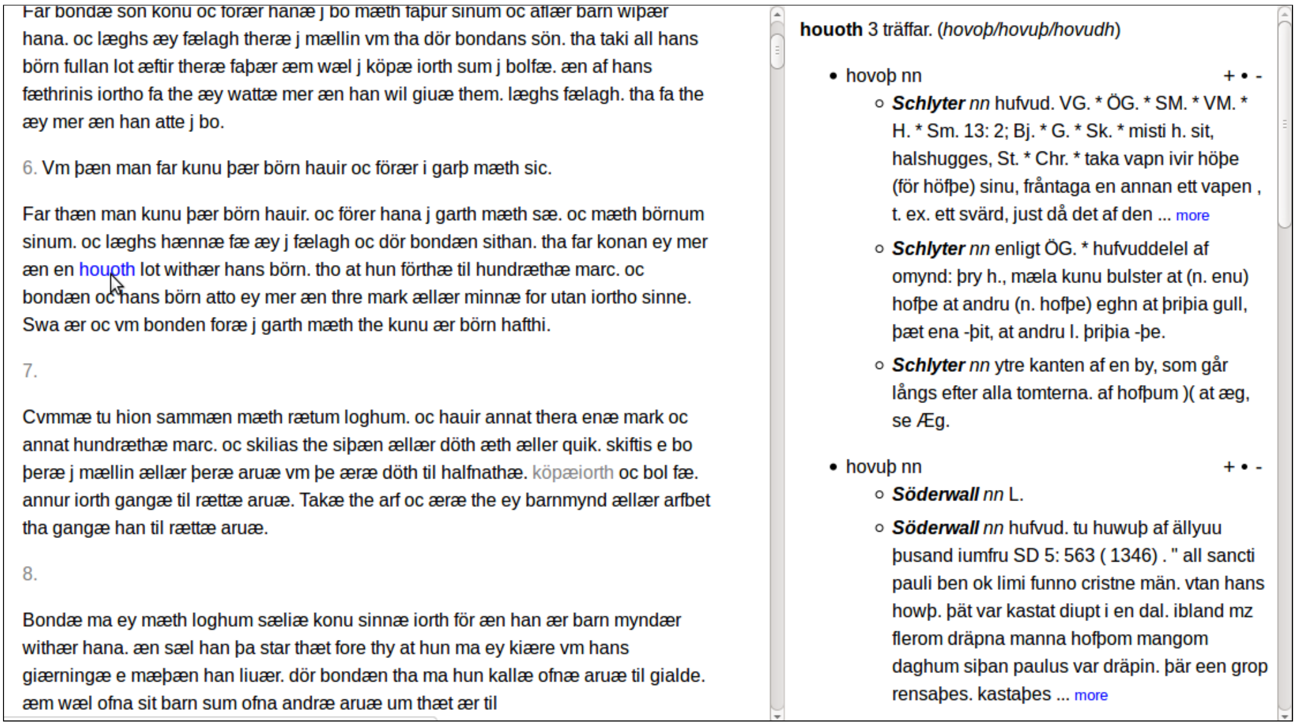
To facilitate the lexical link-up of Old Swedish text, we have further developed methods to efficiently apply minimum weighted edit distance matching to large corpora (Ahlberg and Bouma, Submitted). Depending on the number of matches desired, linking the 3 million token corpus to the 54 000 entry dictionary takes from a few hours up to a day.

The resulting annotation is useful in several ways. First, dictionary entries can be type identifiers to facilitate statistical processing. Secondly, we gain access to the (partial) information about part-of-speech provided in the lexica. Finally, from a user-interface perspective, the link can serve as an assisting technology for studies of the Old Swedish texts.

Regarding this last point, we have created a web interface to display the links between tokens in running text and the lexical entries found by the module. Users can explore the preprocessed texts by clicking on words to retrieve the (hopefully correct) dictionary definitions. A screenshot is shown in Figure 1. Here we see a fragment of Skånelagen (Scania Law, Holm B 76). The user has selected *houoth* 'head' in §6, which is matched in all three dictionaries, with a different spelling in each. The suggested entries can be marked as correct or as errors by the user, by clicking on the $+$ or $-$ symbol to the right of each main entry. The information is saved to a database. Words that do not have

[1] http://spraakbanken.gu.se

[2] http://project2.sol.lu.se/fornsvenska/



Far bondæ son konu oc foræer hanæ j bo mæth faþur sinum oc aflær barn wiþær hana. oc læghs æy fælagh theræ j mællin vm tha dör bondans sön. tha taki all hans börn fullan lot æftir theræ faþær æm wæl j köpæ iorth sum j bolfæ. æn af hans fæthrinis iortho fa the æy wattæ mer æn han wil giuæ them. læghs fælagh. tha fa the æy mer æn han atte j bo.

6. Vm þæn man far kunu þær börn hauir oc föræer i garþ mæth sic.

Far thæn man kunu þær börn hauir. oc fórer hana j garth mæth sæ. oc mæth börnum sinum. oc læghs hænnæ fæ æy j fælagh oc dör bondæn sithan. tha far konan ey mer æn en houoth lot withær hans börn. tho at hun förthæ til hundræthæ marc. oc bondæn oc hans börn atto ey mer æn thre mark ællær minnæ for utan iortho sinne. Swa ær oc vm bonden foræ j garth mæth the kunu ær börn hafthi.

7.

Cvmmæ tu hion sammæn mæth rætum loghum. oc hauir annat thera enæ mark oc annat hundræthæ marc. oc skilias the siþæn ællær döth æth æller quik. skiftis e bo þeræ j mællin ællær þeræ aruæ vm þe æræ döth til halfnathæ. köpæiorth oc bol fæ. annur iorth gangæ til ræettæ aruæ. Takæ the arf oc æræ the ey barnmynd ællær arfbet tha gangæ han til ræettæ aruæ.

8.

Bondæ ma ey mæth loghum sæliæ konu sinnæ iorth för æn han ær barn myndær withær hana. æn sæl han þa star thæt fore thy at hun ma ey kiære vm hans giærningæ e mæþæn han liuær. dör bondæn tha ma hun kallæ ofnæ aruæ til gialde. æm wæl ofna sit barn sum ofna andræ aruæ um thæt ær til

houoth 3 träffar. (*hovoþ/hovuþ/hovudh*)

- hovoþ nn + • -
  - ***Schlyter*** *nn* hufvud. VG. * ÖG. * SM. * VM. * H. * Sm. 13: 2; Bj. * G. * Sk. * misti h. sit, halshugges, St. * Chr. * taka vapn ivir höþe (för höfþe) sinu, fråntaga en annan ett vapen , t. ex. ett svärd, just då det af den ... more
  - ***Schlyter*** *nn* enligt ÖG. * hufvuddelel af omynd: þry h., mæla kunu bulster at (n. enu) hofþe at andru (n. hofþe) eghn at þriþia gull, þæt ena -þit, at andru l. þriþia -þe.
  - ***Schlyter*** *nn* ytre kanten af en by, som går långs efter alla tomterna. af hofþum)(at æg, se Æg.
- hovuþ nn + • -
  - ***Söderwall*** *nn* L.
  - ***Söderwall*** *nn* hufvud. tu huwuþ af ällyuu þusand iumfru SD 5: 563 (1346) . " all sancti pauli ben ok limi funno cristne män. vtan hans howþ. þät var kastat diupt i en dal. ibland mz flerom dräpna manna hofþom mangom daghum siþan paulus var dräpin. þär een grop rensaþes. kastaþes ... more

Figure 1: A screenshot of a linked-up version of Skånelagen.

matches within a predefined threshold are marked in grey, as for instance *köpæiorth* in §7.

3. Future work

In the near future we plan to include the existing computational morphology for Old Swedish in our approach. Our spelling variation module to a large degree also captures inflection, but this needs to be treated consistently.

Our preliminary evaluation also revealed the necessity of handling multi-word tokens and compounding. For instance, *köpæ iorth* 'bought land' cannot currently be matched in the lexicon because we only consider graphic words. Conversely, to link the compound *villhonnugh* 'wild honey', we would have to match substrings to **vilder.ADJ** 'wild' and **hunagh.N** 'honey', respectively.

Finally, a larger annotated corpus is needed for further development and proper evaluation of our methods. Our browsing tool will support this annotation process.

4. References

Yvonne Adesam, Malin Ahlberg, and Gerlof Bouma. 2012. *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...* Towards lexical link-up for a corpus of Old Swedish. In Jancsary, editor, *Empirical Methods in Natural Language Processing: Proceedings of KONVENS 2012 (LThist 2012 workshop)*, page 365–369, Vienna.

Malin Ahlberg and Gerlof Bouma. Submitted. Best first anagram hashing filters for efficient weighted edit distance calculation.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the RANLP Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42. Hissar, Bulgaria.

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech, Marocco. ELRA.

Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010. Diabase: Towards a diachronic BLARK in support of historical studies. In Calzolari, Choukri, Maegaard, Mariani, Odijk, Piperidis, Rosner, and Tapias, editors, *Proceedings of the LREC 2010 workshop on Semantic relations. Theory and Applications*, Valletta, Malta. ELRA.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In Calzolari, Choukri, Declerck, Doğan, Maegaard, Mariani, Odijk, and Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA.

Bryan Jurish. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77, Uppsala, Sweden. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Parsing the past - identification of verb constructions in historical text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Avignon, France.

Carl Johan Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar*, volume 13 of *Saml. af Sweriges Gamla Lagar*. Lund, Sweden.

Knut Fredrik Söderwall. 1884/1953. *Ordbok Öfver svenska medeltids-språket. Ordbok Öfver svenska medeltids-språket. Supplement*. Lund, Sweden.

Predicting alignment performance

Lars Ahrenberg

Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

We present a planned project aimed at evaluating the performance of different word aligners under different conditions. We want to be able to choose the best word aligner for a given corpus and purpose, and, more specifically, to predict a word aligner's performance on the basis of properties of the data. Some of these properties relate to the languages used, while others concern the nature of the parallelism. Performance will be measured wrt both token matches and matches of translation units.

1. Background

It is not uncommon that the output of a system based on statistical learning, such as Google Translate or Giza++ makes you disappointed. This is often due to a mismatch between the data given to the system and its models. Even for a researcher, the success or failure of a given task can be hard to predict. In the case of word alignment there is a rich smorgasbord of systems to choose from, including Giza++ (Och and Ney, 2003), Berkeley Aligner (Liang et al., 2006), Uplug clue aligner (Tiedemann, 2003) and more. Furthermore, each system has a number of parameters that can be set by the user, making the decision even more complex.

2. Goals

The overall goal of the project is to get a better understanding of the adequacy of different word aligners for different types of alignment tasks. In particular, we want answers to questions such as the following:

Given a corpus that we wish to word align, which aligner should we choose? First, the purpose of the alignment is of course an important factor. If the purpose is statistical machine translation, the resulting word alignment will be used by another process that creates data for the decoder, and a common understanding is that we should then go for high recall. However, if the purpose is to build a word aligned resource where translation phenomena can be looked up, high precision should not be sacrificed for better recall. Another relevant aspect is the availability of resources. With millions of aligned sentence pairs available, it is a good choice to use word aligners that employ statistical models, but if the corpus is limited to a few thousand sentence pairs, these may not produce satisfactory results. Similarly, if we have a dictionary, we should quite obviously look for a system that can make good use of it.

A third aspect, which is the one in focus in the project, is the *corpus features*. There are a number of features that are known to affect word alignment systems negatively, such as differences in morphological complexity of the languages concerned, the occurrence of non-local reorderings and null matches. But exactly how such factors affect the outcome is less known. This brings us to the next question:

How can we explain the performance of a word aligner on a given parallel corpus? The general answer to this question is to be found in the fit (or lack thereof) between the features of the corpus and the alignment models used by the system. To give a more useful answer, we need to provide a detailed account of the corpus features and relate them to the system models. I call such a detailed account of corpus features an *alignment profile*. An alignment profile can be defined as a set of probability mass functions that show how token matches are distributed over types. See Table 1 for some examples.

If we know the constraints on alignments that a system assumes, we can be definite about what it can not find. But we cannot know that it will find all instances of what it is looking for. This calls for more empirical studies and brings us to a third question:

How well can we predict the performance of a word aligner, as measured by precision, recall or error rate, from an alignment profile of the test data? (Birch et al., 2008) studied the effects of differences in linguistic structure on machine translation performance, using three dimensions of difference: reordering, morphological complexity, and language relatedness. They concluded that each of these three factors has a significant effect on translation performance as measured by the BLEU score, and a combined model could account for 75% of the variability of the performance for 110 language pairs. They did not report figures on alignment per se, however. But, arguably, word alignment performance should be possible to predict equally well.

A problem, though, is that there exist more reference data for translation than for word alignment. To handle that problem we must use artifical data.

Type description	Examples
Number of tokens	0-1, 1-1, 2-1, m-m, ...
Token positions	same, close, far, ...
Corpus frequency	1, 2, 6-9, ... 100+

Table 1: Dimensions of typing for translation units.

3. Method

We want to test different word aligners with data that vary in alignment profiles. Given an inventory of primitive types, as in Table 1, we can go on to define complex, descriptive properties of alignments in terms of the basic types. For example, we may define *neatness* as the percentage of units that are 1-1, and *faithfulness* as the percentage of all tokens that have received a non-null match.

Language-related properties such as differences in morphological complexity and lexical similarity can also be studied. (Birch et al., 2008) found that differences in type-token ratios, a metric that reflects both these causes, accounted for about a third of the variation in translation performance as measured byn BLEU.

3.1 Metrics

Research on word alignment has mostly been performed in the context of statistical machine translation (SMT) and been evaluated on the basis of machine translation performance. (Och and Ney, 2003) also evaluated intrinsically using Alignment Error Rate (AER) as their measure. This metric has been criticized for good reasons, and with our goals in mind, the major drawback is that it is too coarse and does not reveal qualitative differences. Other common metrics are precision and recall, usually measured on the set of token matches (or links). (Søgaard and Kuhn, 2009) defined a measure they called Translation Unit Error (TUER) which assumes that the alignment is complete and unit-based. This means that there is a decision for all tokens, conforming to the constraint that if tokens i, i', j, j', if $< i, j >$, $< i, j' >$, and $< i', j >$ are aligned, then so is $< i', j' >$. Metrics based on translation units are actually more relevant for purposes of resource creation and will be used in the project.

3.2 Corpus generation

Available natural gold standards can be used, where available, but to systematically study the effects of different alignment profiles, we need to be able to generate data with known properties. For this purpose we use probabilistic synchronous grammars.

The synchronous grammars generate sentence pairs with their word alignments. Null alignments as well as many-to-many alignments (including many-to-one and one-to-many) can be generated and the frequency of these alignments is determined by the probabilities assigned to rules that define them. Similarly, the amount of reorderings in the corpus is determined by the probabilities of the rules that have reordered constituents. Some rule examples are illustrated in Figure 1.

The vocabulary is divided into parts-of-speech with a different treatment of function words and content words. Each content word, for both source and target vocabularies, is associated with one or more *cepts* where a cept represents a meaning. The cepts determine possible alignments. Multi-word expressions start life as single tokens in the grammar and are then split in a separate process to produce many-to-many alignments.

The current grammars have rules of depth one and are thus not expressive enough to be able to generate all types of alignment phenomena that occur in real translations (Søgaard and Kuhn, 2009). Still, the types of alignments they can generate allow for a wide range of alignment profiles.

NP → N, N, 1-1, 0.46
NP → N, P N, 0-1 1-2, 0.10
NP → AP N, N AP, 1-2, 2-1, 0.10
A: 500, 100, 50, 0.06, 0.04
N: 4000, 500, 100, 0.05, 0.04
P: 20, 20, 10, 0.16, 0.12

Figure 1: Examples of rules and vocabulary definitions. Positive numbers indicate the number of lexical items to be generated with one, two or three meanings. Numbers between 0 and 1 are probabilities.

Scores	**Gold**	**IBM-1**	**IBM-2**	**IBM-4**
Precision	1	0.843	0.888	0.929
Recall	1	0.785	0.831	0.858
Faithfulness	0.963	0.916	0.949	0.929
Neatness	0.807	0.780	0.840	0.886

Table 2: Scores and profiles for three alignment models compared with a gold standard.

In Table 2 we show data from a run of Giza++ on an artificial corpus with 50,000 sentence pairs. Sentences varied in length between 2 and 100 tokens with an average of 10.4. The vocabulary defined by the grammar had just under 7000 source stems and some 8400 target stems. The morphology was simple with an equal number of inflections for both languages. As is expected precision and recall improve with better models, but all systems' alignment profiles differ from the that of the gold standard. In particular, we can see that IBM-4, which has by far the best scores in terms of precision and recall, exaggerates the neatness and underestimates the faithfulness of this corpus.

4. References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings on Empirical Methods in Natural Language Processing (EMNLP)*, pages 745–754, Honolulu, USA, October.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*, pages 104–111, New York City, USA, June.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of SSST-3 Third Workshop on Syntax and Structure in Statistical Translation*, pages 19–27, Boulder, Colorado, June.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the Tenth Conference of the EACL*, pages 339–346.

Extractive document summarization - an unsupervised approach

Jonatan Bengtsson*, **Christoffer Skeppstedt**†, **Svetoslav Marinov***

*Findwise AB, †Tickster AB, Gothenburg, Sweden
*{jonatan.bengtsson,svetoslav.marinov}@findwise.com, †christoffer@tickster.com

Abstract

In this paper we present and evaluate a system for automatic extractive document summarization. We employ three different unsupervised algorithms for sentence ranking - TextRank, K-means clustering and previously unexplored in this field, one-class SVM. By adding language and domain specific boosting we achieve state-of-the-art performance for English measured in ROUGE Ngram(1,1) score on the DUC 2002 dataset - 0,4797. In addition, the system can be used for both single and multi document summarization. We also present results for Swedish, based on a new corpus - featured Wikipedia articles.

1. Introduction

An extractive summarization system tries to identify the most relevant sentences in an input document (aka single document summarization, SDS) or cluster of similar documents (aka multi-document summarization, MDS) and uses these to create a summary (Nenkova and McKeown, 2011). This task can be divided into four subtask: document processing, sentence ranking, sentence selection and sentence ordering. Section 2. describes all necessary document processing.

We have chosen to work entirely with unsupervised machine learning algorithms to achieve maximum domain independence. We utilize three algorithms for sentence ranking - TextRank (Mihalcea and Tarau, 2004), K-means clustering (García-Hernández et al., 2008) and One-class Support Vector Machines (oSVM) (Schölkopf et al., 2001). In three of the summarization subtasks, a crucial component is the calculation of sentence similarities. Both the ranking algorithms and the similarity measures are presented in more detail in Section 3.

Sections 4. and 5. deal with the different customizations of the system to tackle the tasks of SDS and MDS respectively. Here we also describe the subtasks of sentence selections and ordering.

In Section 6. we present the results from the evaluation on Swedish and English corpora.

2. Document processing

The minimal preprocessing required by an extractive document summarization system is sentence splitting. While this is sufficient to create a baseline, its performance will be suboptimal. Further linguistic processing is central to optimizing the system. The basic requirements are tokenization, stemming and part-of-speech (POS) tagging. Given a language where we have all the basic resources, our system will be able to produce a summary of a document.

Sentence splitting, tokenization and POS tagging are done using OpenNLP (http://opennlp.apache.org) for both English and Swedish. In addition we have explored several other means of linguistic analysis such as Named Entity Recognition (NER), keyword extraction, dependency parsing and noun phrase (NP) chunking. Finally we can augment the importance of words by calculating their *term-frequency* times *inverse document frequency* (*TF*IDF*) score.

NER for English is performed with Stanford Named Entity Recognizer (http://nlp.stanford.edu/software) while for Swedish we use OpenNLP. The latter library is also used for NP chunking for English. Dependency parsing is performed by MaltParser (http://maltparser.org).

3. Sentence Ranking

A central component in an extractive summarization system is the sentence ranking algorithm. Its role is to assign a real value rank to each input sentence or order the sentences according to their relevance.

3.1 TextRank

TextRank (Mihalcea and Tarau, 2004) models a document as a graph, where nodes corresponds to sentences from the document, and edges carry a weight describing the similarity between the nodes. Once the graph has been constructed the nodes are ranked by an iterative algorithm based on Google's PageRank (Brin and Page, 1998).

The notion of sentence similarity is crucial to TextRank. Mihalcea and Tarau (2004) use the following similarity measure: $Similarity(S_i, S_j) = \frac{|S_i \cap S_j|}{log|S_i|+log|S_j|}$, where S_i is a sentence from the document, $|S_i|$ is its length in words and $|S_i \cap S_j|$ is the word overlap between S_i and S_j.

We have tested several other enhanced approaches, such as *cosine*, *TF*IDF*, *POS tags* and *dependency tree* based similarity measures.

3.2 K-means clustering

García-Hernández et al. (2008) adapt the well-known K-means clustering algorithm to the task of document summarization. We use the same approach and divide sentences into k clusters from which we then select the most salient ones. We have tested three different ways for sentence relevance ordering - position, centroid and TextRank-based. The value of k is conditioned on the mean sentence length in a document and the desired summary length.

Each sentence is converted to a word vector before the clustering begins. The vectors can contain all unique words of the document or a subset based on POS tag, document keywords or named entities.

3.3 One-class SVM

oSVM is a previously unexplored approach when it comes to unsupervised, extractive summarization. Similarly to the K-means algorithm, sentences are seen as points in a coordinate system, but the task is to find the outline boundary, i.e. the support vectors that enclose all points. These vectors (or sentences) arguably define the document and are therefore interesting from a summarization point of view.

For the kernel function we choose the sentence similarity measures (cf. 3.1). Similarly to choosing k in (3.2), the number of support vectors is dependent on the mean sentence and desired summary lengths.

4. Single document summarization

For SDS we can use domain specific knowledge in order to boost the sentence rank and thus improve the performance of the system. As an example, in the domain of newspaper articles the sentence position tends to have a significant role, with initital sentences containing the gist of the article. We use an inverse square function to update the sentence ranks: $Boost(S_i) = S_i.rank * (1 + \frac{1}{\sqrt{S_i.pos}})$, where $S_i.rank$ is the prior value and $S_i.pos$ is the position of the sentence in the document. We see such boosting functions as important steps for domain customization.

Once the sentences have been ranked the selection and ordering tasks are relatively straightforward - we take the highest ranked sentences until a word limit is reached and order these according to their original position in the text.

5. Multi document summarization

When it comes to MDS, two different approaches have been tested. The first one is to summarize a document cluster by taking all sentences in it. The other approach is based on the work of (Mihalcea and Tarau, 2005) who use a two stage approach, where each document is first summarized and then we summarize only the summaries (2-stageSum).

MDS shares the same ranking algorithms as SDS, coupled with specific sentence selection and ordering. We rely on similarity measures (cf. 3.1) to avoid selecting near sentence duplicates and adopt a topic/publication date-based approach for sentence ordering (Bollegala et al., 2006).

6. Evaluation

The system is evaluated on the DUC 2002 corpus, which consists of 567 English news articles in 59 clusters paired with 100 word summaries. For Swedish we use a corpus of 251 featured Wikipedia articles from 2010, where the introduction is considered to be the summary.

We rely on the ROUGE toolkit to evaluate the automatically generated summaries and use Ngram(1,1) F_1 settings, as these have been shown to closely relate to human ratings (Lin and Hovy, 2003), without stemming and stop word removal. Two kinds of baseline systems are also tested - random selection and leading sentence selection (see Table 1).

	English		Swedish
Algorithm	SDS	MDS	SDS
TextRank (*TF*IDF, POS*)	**0.4797**	0.2537	**0.3593**
K-means (*TF*IDF, POS*)	0.4680	0.2400	0.3539
oSVM (*cosine*)	0.4343		0.3399
2-stageSum (*TextRank-based*)		**0.2561**	
(Mihalcea and Tarau, 2004)	0.4708		
(García-Hernández et al., 2008)	0.4791		
Baseline$_{lead}$	0.4649	0.2317	0.3350
Baseline$_{rand}$	0.3998	0.2054	0.3293

Table 1: Results and Comparison

7. Conclusion

In this paper we have presented a system capable of doing both SDS and MDS. By relying on unsupervised machine learning algorithms we achieve domain independence. With relatively little language dependent processing the system can be ported to new languages and domains. We have evaluated three different algorithms for sentence ranking where oSVM is previously unexplored in this field. By adding domain knowledge in the form of sentence rank boosting with the TextRank algorithm we receive higher ROUGE scores than other systems tested on DUC 2002 dataset. In addition, we have tested the system for Swedish on a new corpus with promising results.

8. References

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2006. A bottom-up approach to sentence ordering for multi-document summarization. In *In Proceedings of the COLING/ACL*, pages 385–392.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April.

René Arnulfo García-Hernández, Romyna Montiel, Yulia Ledeneva, Eréndira Rendón, Alexander Gelbukh, and Rafael Cruz. 2008. Text Summarization by Sentence Extraction Using Unsupervised Learning. In *Proceedings of the 7th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*, MICAI '08, pages 133–143, Berlin, Heidelberg. Springer-Verlag.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1 of *NAACL '03*, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Rada Mihalcea and Paul Tarau. 2005. Multi-document summarization with iterative graph-based algorithms. In *in Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*.

Ani Nenkova and Kathleen McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July.

Small versus closed syntactical patterns in authorship attribution

Johanna Björklund, Jonas Lindh Morén

Dept. Computing Science
Umeå University
johanna@cs.umu.se, dv08jmh@cs.umu.se

Abstract

We report on an experimental study that compares the usefulness of small and closed syntactical patterns for authorship attribution. The classification task at hand consists in distinguishing between the writings of 5 – 20 different authors, drawn from a corpus of blog and news feeds. The results suggest that small patterns outperform closed patterns both in terms of accuracy and efficiency when the number of authors increase.

1. Introduction

The authorship attribution problem consists in the following: Given a set of authors, samples of their work, and an anonymous document d, decide who among the authors wrote d. Algorithmic solutions to this problem could help detect fraud, plagiarism, and internet propaganda, and could also aid literary analysis and collaborative writing.

Since the problem is in essence an inference problem, it is typically approached by (i) adopting either a predefined set of features or mining such a set from the training data, (ii) training a classification model through, e.g., latent semantic analysis or naive Bayes, and (iii) applying the trained model to classify anonymous documents.

A multitude of features have been considered in literature, ranging from sentence length and the use of punctuation to functional-words and syntactical patterns (Stamatatos, 2009). The latter two arguably depend less on the subject matter than vocabulary analysis, which is an advantage when the training data is heterogeneous in terms of topic, context, and genre.

The rationale behind syntactical patterns is that just as an author prefers some words to others, so does he or she tend towards certain grammatical constructions. The patterns considered can be part-of-speech tags, connected fragments of syntax trees, or more complex structures. In Figure 1, a repeated pattern in a syntax tree has been highlighted.

For practical purposes, syntactical patterns are typically included in a battery of features, assembled to capture various aspects of language such as vocabulary, grammar, and common misspellings. For research purpose however, it makes sense to study alternative forms of syntactical patterns in isolation, so as to better understand how they contribute to the classification process.

In this work, we compare the usefulness of two types of syntactical patterns for authorship attribution. The first of these are fragments of syntax-trees of height one, i.e., single nodes together with their immediate children. An advantage of such patterns, which we simply call *small patterns*, is that they are easy to mine from training data.

The second type of pattern was introduced by Chi et al. (2005) under the name *closed and frequent subtrees*. A syntax tree t is said to contain a pattern tree s if s occurs as a connected subgraph of t. Let D be a set of syntax trees. The *support* of a pattern s in D is the cardinality of the set $\{t \in D \mid t \text{ contains } s\}$. A pattern tree t is *closed* if no proper supertree of t has the same support as t. For example, if t has support 4, and no supertree of t has a support larger than 3, then t is closed. If a pattern is not closed, then it is contained in a strictly larger but closed pattern that occurs at the same places in the data, so it suffices to include this larger pattern in the feature set.

Chi et al. (2005) gave an algorithm for mining closed patterns from training data. The algorithm employs various pruning techniques to avoid unnecessary computation. The algorithm was later adopted by Kim et al. (2010) who studied closed patterns as a feature for authorship attribution. Their experiments were run on data sets consisting of news articles and movie reviews. In the case of binary classification, when the task is to separate the works of two authors, Kim et al. found closed patterns to yield better performance than small patterns. The average accuracies for small and closed patterns for news articles were 89.8 versus 95.3, and for movie reviews 92.9 versus 95.8.

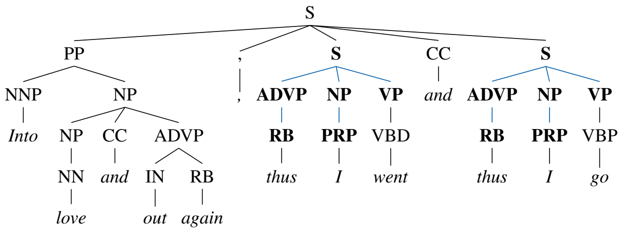


Figure 1: A reoccurring pattern in a syntax tree. The excerpt is from the poem *Theory* by Dorothy Parker.

2. Experiments

To better understand the relative strengths and weaknesses of small and closed patterns, we conduct a series of comparative experiments. These are run on the ICWSM corpus which consists of nearly 200 gigabyte of XML-formatted blog and news feeds. ICWSM is organised as a sequence of entries together with various degrees of meta information, usually containing Author, Language, and Web Resource. The corpus was collected by Tailrank (now Spinn3r.com) during the period August – October 2008 and made available to the scientific community in 2009 (ICWSM, 2009).

The experimental setup is as follows. For the number of authors $n \in [5, 20]$ and the index $i \in [1, 70]$, a sample of n authors is drawn from the ICWSM corpus, and for each author 10 of his or her blog entries. The chosen documents are divided at random into a test set T_n^i containing 10 entries, and a training set S_n^i of size $10n - 10$.

In experiment $(n, i) \in [5, 20] \times [1, 70]$, a pair of k-nearest neighbour (KNN) classifiers A and B are trained on S_n^i using small and closed patterns, respectively. The two classifiers are then set to categorize the documents in T_n^i. Each classifier ranks the candidate authors in order of likelihood, and their performance is scored according to the function $1/(2^{r-1})$, where r is the rank of the correct author. A classifier is thus given the score 1 if it ranks the correct author the highest, and 0.5 if the correct author is its second guess. For each number of authors n, we thereby obtain 70 pairs of values, each pair telling us how successful A and B were, on average, in one of the experiments.

The experiments required two weeks' worth of computation time on a modern stationary computer. Approximately 80 % of the work was spent on mining closed patterns, and the remainder on mining small patterns, training the classifiers, and computing the test results. It stands to reason that the more complex closed patterns might be at a better advantage if the data sets were greater, but at the same time, it is the cost of mining closed patterns that prevents us from conducting larger experiments.

2.1 The KNN classifiers

Simply put, the KNN approach consists in identifying documents with vectors in a feature space and then classifying the objects in the test set based on their k nearest neighbours in the training set. Two vectors are considered close (or 'similar') if their dot product is large. We combine KNN with latent semantic analysis, in which the feature space is projected onto a reduced space before classification starts. This smaller space is derived through singular value decomposition, in such a way that documents that were close in the original space remain close in the reduced space. This adds work to the training phase, but it makes the classification faster and lessens the risk of over-fitting.

In the KNN classification, we took k to be 10 and projected the vector space onto a reduced space with as many dimensions as the number of authors in the particular experiment. These parameter values were obtained through a range of preliminary tests and are favorable for both types of syntactical patterns.

3. Results and analysis

The outcome of the experiments is presented in Figure 2. The line marked with triangles indicates the scores for small patterns, and the line with crosses the scores for closed patterns. The scores for function words and random guesses are included for reference (the latter of these was statistically computed, hence the smooth curve). In comparison with the sample texts of Kim et al. (2010), the samples derivable from ICWSM are restricted, heterogeneous, and noisy. In combination with a larger number of author candidates, this leads to lower performance scores.

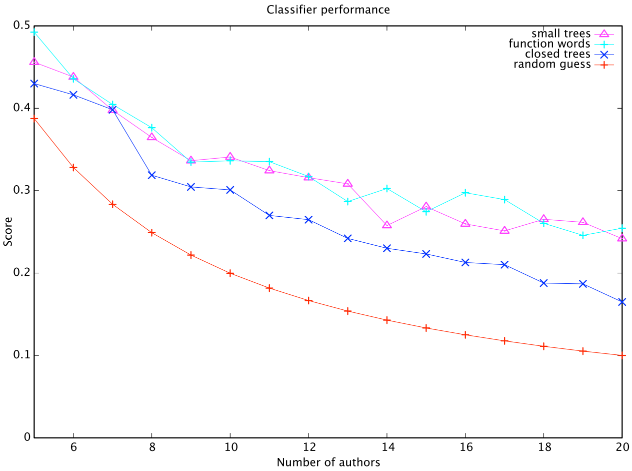


Figure 2: A comparative plot of the performances of a pair of KNN classifiers using small patterns and closed patterns as features, respectively. The vertical axis indicates the performance score; the horizontal axis the number of authors.

The results suggest that small patterns outperform closed patterns when the number of authors increase. When applying the Wilcoxon sign-ranked test with $\alpha = 0.05$, we can reject the null hypothesis that closed trees are at least as good as small trees whenever we have more than 7 authors. Since it is computationally more expensive to mine closed rather than small patterns, small patterns appears to be the better choice when the number of authors is large, but the available text sample for each author is small.

On the other hand, closed patterns may be valuable when the classification model is meant to guide human analysis. For example, in Stylistics, it is likely more interesting to learn that a certain author often employs a particular, complex, syntactical construction, rather than being informed that his or her usage of a series of syntactical fragments can be described through a certain linear equation.

Acknowledgments. We are grateful to Lars Bergström at the University of Nottingham for his contributions to the software framework underlying our experiments.

4. References

Y. Chi, Y. Xia, Y. Yang, and R. R. Muntz. 2005. Mining closed and maximal frequent subtrees from databases of labeled rooted trees. *IEEE Transactions on Knowledge and Data Engineering*, 17:2001.

S. Kim, H. Kim, T. Weninger, and J. Han. 2010. Authorship classification: a syntactic tree mining approach. In *Proceedings of the ACM SIGKDD Workshop on Useful Patterns*, pages 65–73, New York, NY, USA. ACM.

E. Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

ICWSM. 2009. 3rd international AAAI conference on weblogs and social media. http://www.icwsm.org/2009/, accessed June 29th, 2010.

HMM based speech synthesis system for Swedish Language

Bajibabu Bollepalli, Jonas Beskow, Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Stockholm
{bajibabu,beskow}@kth.se, jocke@speech.kth.se

Abstract

This paper describes the Hidden markov models-based Text-to-Speech system (HTS) for Swedish language. In this system, speech is represented by a set of parameters. This scheme of representation enables us to modify these parameters, and reconstruct speech from them. Due to this flexibility, it is possible to generate desired voice qualities, e.g. breathy voice, creaky voice, etc.. Here, we describe the architecture of the current state-of-the-art HTS system and present preliminary evaluation results for Swedish language.

1. Introduction

The aim of a speech synthesis system is to generate a human like voice from arbitrary text. These systems have been under development for several decades. Recent progress in speech synthesis has produced synthesizers with high intelligibility and naturalness. Traditional concatenative or unit-selection based speech synthesis systems (A. Hunt and A. Black 1996), which synthesize the speech by joining different length speech-units (like phones, diphones and syllables, etc.) derived from the natural speech, requires large amount of training data to synthesis good quality of speech. However, it is very difficult to collect and store a large speech corpora. Furthermore, the quality of synthesis in these systems depend upon the goodness in joining of the natural speech-units. To overcome these problems, Hidden markov model (HMM) based speech synthesis system (HTS) was proposed by T. Yoshimura et. al (1999). In HTS, speech is represented by spectral, excitation and durational parameters. These parameters are modeled by contextual-dependent HMMs. Due to this parametric representation, it has following advantages:

1. Smooth and natural sounding speech can be synthesized from small amount of speech corpora.
2. The voice characteristics can be changed.

In this paper, we train a HTS system to synthesis speech in Swedish language.

2. Overview of a basic HTS system

Figure 1 shows an architecture of a basic HMM-based speech synthesis system (H. Zen et. al, 2005). A HTS system mainly consists of two parts: 1) training and 2) synthesis

2.1 Training

Extraction of parameters:
The training part consists of extracting parameters from a given speech database and modeling these parameters by contextual HMMs. To extract features from a given speech signal we used the a high quality analysis tool called STRAIGHT (H. Kawahara et. al 1999). This system extracts three kinds of parameters: 1) spectral, 2) excitation, and 3) aperiodic measures from a speech signal. In this work, we used the Mel-cepstral coefficients (MCEPs) and fundamental frequency (F0) as spectral and excitational parameters respectively. Figure 2 shows the procedure to extract MCEPs from a speech signal.

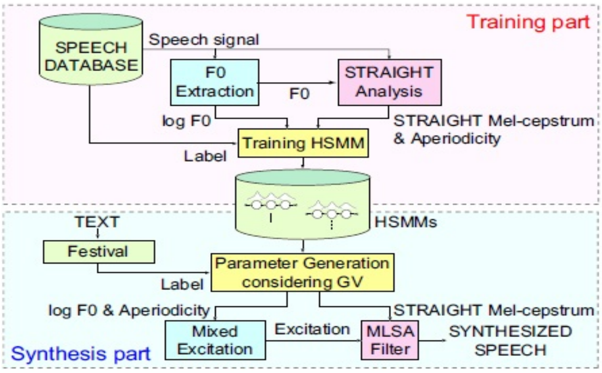


Figure 1: Architecture of a HMM-based speech synthesis system (adopted from H. Zen et. al, 2005).

To extract the fundamental frequency (F0), we performed the voting between the outputs of 1)an instantaneous-frequency-amplitude-spectrum based algorithm, 2)a fixed point analysis called TEMPO, and 3)the ESPS get-F0 tool. The aperiodicity measures were calculated as the ratio between the lower and upper smooth spectral envelopes, and averaged across five frequency subbands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz).

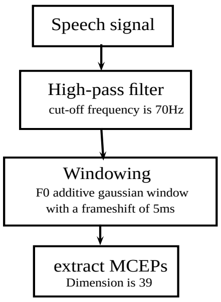


Figure 2: Extraction of spectral parameters from speech signal

HMM training:

We followed the same training procedure showed in (H. Zen et. al, 2005). In the training, all extracted parameters and durations are modeled in single unified framework. For this, multi-stream model structure is used for simultaneous and synchronous modeling of extracted parameters. We used the Hidden-semi markov model (HSMM), which is same as HMMs except it estimates the state duration probabilities in each iteration of the training. As fundamental frequency (F0) values consist of continuous values in voice regions and discrete symbols (zero) in unvoiced regions, a Multi-space density (MSD) distribution is used to model statistics of F0. The following contextual information is considered in the training to build contextual-HMMs.

- phoneme:
  - {preceding, current, succeeding} phoneme
  - position of current phoneme in current syllable
- syllable:
  - number of phonemes at {preceding, current, succeeding} syllable
  - accent and stress of {preceding, current, succeeding} syllable
  - position of current syllable in current word
  - number of {preceding, succeeding} accented and stressed syllables in current phrase
  - number of syllables {from previous, to next} accented and stressed syllable
  - vowel within current syllable
- word:
  - guess at part of speech of {preceding, current, succeeding} word
  - number of syllables in {preceding, current, succeeding} word
  - position of current word in current phrase
  - number of {preceding, succeeding} content words in current phrase
  - number of words {from previous, to next} content word
- phrase:
  - number of syllables in {preceding, current, succeeding} phrase
  - position in major phrase
  - ToBI endtone of current phrase
- utterance:
  - number of syllables in current utterance

2.2 Synthesis

In the synthesis part, first a given text to be synthesized is converted to a context-dependent label sentence and a sentence MSD-HSMM is constructed by concatenating the parameter-tied context-dependent MSD-HSMMs. Secondly, state durations maximizing their probabilities are determined. Thirdly, speech parameters are generated by using speech parameter generation algorithm. Finally, speech waveform is synthesized directly from the generated MCEPs, F0 and aperiodicity measure sequences using the STRAIGHT vocoder.

Table 1: MOS-based evaluation of the HMM-based speech synthesis system

System	Naturalness	Intelligibility
HTS	3.8	4.2

3. Experimental evaluation

We used 1000 Swedish sentences for training with an average duration of 3 seconds. The contextual-labels were generated by using RULSYS, which is developed by KTH and has a long history (Carlson et. al, 1982). Five state left-to-right MSD-HSMMs without skip paths were used for training. Each state has a single Gaussian probability distribution function (pdf) with a diagonal covariance matrix as the state output pdf and a single Gaussian pdf with a scalar variance as the state duration pdf.

A perceptual evaluation was conducted to asses the performance of the system. Evaluation is based on the Mean opinion score (MOS) concerning the naturalness and the intelligibility on a scale of one to five where one stands for “bad” and five stands for “excellent”. A total of 5 sentences were synthesized by the system. A group of 10 listeners, comprising both speech and non-speech experts were asked to express their opinion for each sentence on a MOS scale. The results of the test are showed in Table 1. The results indicate that the Swedish speech synthesized by the basic HTS system were perceived natural and intelligible.

These results are encouraging since this is a basic HTS system built for the Swedish language and many further improvements are possible. In our future work, we intend to adapt a more sophisticated model for speech reconstruction, a parametric model of voice source to synthesis different voice styles, and include more prosodic properties in order to increase the naturalness of the produced speech.

4. References

A. Hunt, and A. Black (1996) Unit selection in a concatenative speech synthesis system using a large speech database. *in proceedings of ICASSP*

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *in Proceedings of Eurospeech*, pp. 2347-2350.

H. Zen, T. Toda (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. *in proceedings of INTERSPEECH*, Lisbon, pp. 93-96.

H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, vol. 27, pp. 187-207.

Carlson, Granström, and Hunnicutt (1982). A multi-language text-to-speech module. *in proceedings of ICASSP*, vol. 3.

Growing a Swedish constructicon in lexical soil

Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice,
Rudolf Rydstedt, Emma Sköldberg, Sofia Tingsell

Department of Swedish, University of Gothenburg
firstname.lastname@svenska.gu.se

1. Introduction

Large-scale linguistic resources typically consist of a lexicon and/or a grammar, and so do the linguistic components in language technology (LT) applications. Lexical resources mainly account for words, whereas grammars focus on general linguistic rules. Consequently, patterns that are too general to be attributed to individual words but too specific to be considered general rules are peripheral from both perspectives and hence have tended to be neglected. Such constructions are not, however, a trivial phenomenon that can simply be disregarded. On the contrary, semi-productive, partially schematic multi-word units are highly problematic for LT (Sag et al., 2002), L2 acquisition (Prentice and Sköldberg, 2011), and, given that idiosyncracies are typically attributed to the lexicon, lexicography. They are also quite common (cf. e.g., Jackendoff 1997, 156).

In response to this, we are currently building a Swedish *constructicon*, based on principles of Construction Grammar and developed to be an integral component of the tightly interlinked computational lexical resource network developed in the Swedish FrameNet++ project. The constructicon project is a collaboration involving experts on (construction) grammar, LT, lexicography, phraseology, second language research, and semantics.

2. Swedish FrameNet++

The goal of the Swedish FrameNet++ project (Borin et al., 2010) is to create a large-scale, freely available integrated lexical resource for Swedish – so far lacking – to be used as a basic infrastructural component in Swedish LT research and in the development of LT applications for Swedish.[1] The specific objectives of the project are

- integrating a number of existing free lexical resources into a unified lexical resource network;
- creating a full-scale Swedish FrameNet;
- developing methodologies and workflows which make maximal use of LT tools in order to minimize the human effort needed in the work.

3. Constructions

Constructions (cx) are conventionalized pairings of form and meaning, essentially signs, of varying degrees of generality (table 1). A significant subpart of these are semi-general linguistic patterns, neither general rules of grammar nor lexically specific idiosyncracies (partially schematic cx in table 1). Despite being both numerous and common, these patterns have a long history of being neglected. For the last few decades, however, the study of such cx is on the rise, due to the development of Construction Grammar (CG; Goldberg 1995 and others) and other cx-oriented models. Furthermore, cx have also been gaining increased attention from some lexicalist perspectives, especially through Sign-Based CG (Boas and Sag, to appear). Despite this development, there are still few, if any, large-scale cx accounts.

Degree of schematicity	**Examples**
fully schematic	[V NP]$_{\text{VP}}$, [NP VP]$_{\text{S}}$, stem V-Past (e.g. *walk-ed, smell-ed*)
partially schematic	*the* [AdjP] (*the rich/hungry/young*), [time expression] *ago* (*six days/beers ago*)
fully filled and partially flexible	*go*[tense] *postal*, *hit*[tense] *the road*
fully filled and fixed	*blue moon, by and large,* *children, ink, blue*

Table 1: Types of constructions

Parallel to CG, Fillmore (1982) and associates have also developed Frame Semantics, which in turn constitutes the base for FrameNet (Baker et al., 1998). By its historical and theoretical connections to CG, FrameNet is well suited for inclusion of cx patterns. There is also a growing appreciation for the need to do so. Accordingly, an English constructicon is being developed as an addition to the Berkeley FrameNet (Fillmore et al., to appear). In a similar fashion, the Swedish constructicon will be an extension of SweFN (Lyngfelt et al., 2012).

4. The Swedish constructicon

The Swedish constructicon is modeled on Berkeley's English constructicon: The cx are presented with definitions in free text, schematic structural descriptions, definitions of cx elements, and annotated examples.[2] We try to keep the analyses simple, roughly on the level of complexity commonly used for lexical information. Our expectation is that a corresponding level of complexity will work for a constructicon as well. More detailed analyses are labor-intensive and

[1] See <http://spraakbanken.gu.se/eng/swefn/> for more information about the project, including a development version of the Swedish framenet.

[2] The development version of the Swedish constructicon is available online at <http://spraakbanken.gu.se/swe/resurs/konstruktikon/utvecklingsversion>. At the time of writing, it contains a handful of cx in various stages of completion.

therefore difficult to conduct on a large scale. Still, it is necessary to add some complexity compared to lexical definitions, since descriptions of syntactic cx also must contain constituent structure. Therefore, initially the core of the cx descriptions consists of a simple structural sketch and a free text definition of the dictionary type.

A cx type of high initial priority are the partially schematic cx in table 1. These cx are somewhat similar to fixed multi-word expressions and are fairly close to the lexical end of the cx continuum. They should be easier to identify automatically than fully schematic cx, and are therefore a natural initial target for the development of LT tools. Also, these cx are the ones closest at hand for integration into lexical resources.

The constructicon is usage-based: Cx are identified and characterized using Språkbanken's richly annotated corpora (Borin et al., 2012). The linguistic annotations are a vital feature for this project, since a cx may be defined by constraints on different levels: word, word-form, part of speech, morphosyntactic category, grammatical function, intonation, information structure, etc.

In an ongoing pilot study, we have mainly relied on traditional manual linguistic analysis for the identification of cx, but we will also develop tools to identify cx automatically. For this purpose, StringNet (Wible and Tsao, 2010) is one of many possible research directions. StringNet identifies recurring n-gram patterns, where every unit is classified on three levels: word form, lemma, and grammatical category. Filtering these n-grams with methods for automatic morphology induction (Hammarström and Borin, 2011) and word segmentation (Hewlett and Cohen, 2011) should make it possible to identify likely cx candidates.

Developing tools for automatic identification of cx is both a methodological approach and a highly relevant research objective in its own right. If we are able to automatically identify cx in authentic text, the ambiguity that has always plagued automatic syntactic analysis can be greatly reduced. Thus, it has been shown how pre-identification of different types of local continuous syntactic units may improve a subsequent global dependency analysis (e.g. Attardi and Dell'Orletta 2008). Our hypothesis is that cx can be used in the same way, and exploring this would be a valuable contribution to LT research.

5. Summary

In summary, the constructicon is not only a desirable and natural development of the FrameNet tradition, it is also potentially useful in a number of areas, such as LT, lexicography and (L2) acquisition research and teaching. In addition to these practical uses, we hope that this work will lead to theoretically valuable insights about the relation between grammar and lexicon.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Swedish Framenet++*, VR dnr 2010-6013), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, by the European Commission through its support of the META-NORD project under the ICT PSP Programme, grant agreement no 270899, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldberg, sponsored by the Knut and Alice Wallenberg Foundation.

6. References

Giuseppe Attardi and Felice Dell'Orletta. 2008. Chunking and dependency parsing. In *LREC Workshop on Partial Parsing*, pages 27–32, Marrakech.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING 1998*, pages 86–90, Morristown, NJ. ACL.

Hans Boas and Ivan Sag, editors. to appear. *Sign-Based Construction Grammar*. CSLI, Stanford.

Lars Borin, Dana Danélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, Istanbul.

Charles Fillmore, Russell Lee Goldman, and Russell Rhodes. to appear. The FrameNet constructicon. In Hans Boas and Ivan Sag, editors, *Sign-Based Construction Grammar*. CSLI, Stanford.

Charles Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co, Seoul.

Adele Goldberg. 1995. *Constructions*. University of Chicago Press, Chicago & London.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Daniel Hewlett and Paul Cohen. 2011. Word segmentation as general chunking. In *Proceedings of CoNLL 2011*, pages 39–47, Portland, Oregon.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Adding a constructicon to the Swedish resource network of Språkbanken. In *Proceedings of KONVENS 2012*, pages 452–461, Vienna. ÖGAI.

Julia Prentice and Emma Sköldberg. 2011. Figurative word combinations in texts written by adolescents in multilingual school environments. In Roger Källström and Inger Lindberg, editors, *Young urban Swedish. Variation and change in multilingual settings*, pages 195–217. University of Gothenburg, Dept. of Swedish.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multi-word expressions: A pain in the neck for NLP. In *Proceedings of CICLING-2002*.

David Wible and Nai-Lung Tsao. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 25–31, Los Angeles. ACL.

An Evaluation of Post-processing Google Translations with Microsoft® Word

Håkan Burden, Jones Belhaj, Magnus Bergqvist, Joakim Gross,
Kristofer Hansson Aspman, Ali Issa, Kristoffer Morsing, Quishi Wang

Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
burden@chalmers.se

1. Introduction

A common problem with statistical machine translation (SMT) is that the candidate translations are often extra-grammatical. Recent research has tried to improve on the quality of the translations made by SMT systems by post-processing the candidate translations using a grammar checker (Huet et al., 2010; Stymne and Ahrenberg, 2010). The SMT systems in these studies require training on parallel corpora, while most end-users want a general-purpose translator and neither have the necessary knowledge nor a representative corpora for training an SMT system. Thus the popularity of systems such as Google Translate.

2. Evaluation through Replication

We decided to evaluate the performance of Google Translate and the possible improvements on grammatical fluency through post-processing the candidate translations by Microsoft® Word, replicating previous research done by Stymne and Ahrenberg (2010).

2.1 Replicated Study

Stymne and Ahrenberg (2010) evaluate the impact of post-processing SMT candidate translations by first training Moses (Koehn et al., 2007) on 701 157 English-Swedish sentence pairs taken from the EuroParl corpus (Koehn, 2005). The resulting SMT system was then evaluated on 2 000 sentences of EuroParl. The translations were post-processed by Granska (Carlberger et al., 2004), a grammar checker for Swedish. If there were more than one possible correction according to Granska the first was always chosen. The impact of the grammar checker was evaluated by both an automatic analysis using BLEU (Papineni et al., 2002) as well as a manual inspection of the first 100 suggested corrections made by Granska.

2.2 Replication Setup

In our replication we chose to use Google Translate instead of Moses since we wanted an SMT system that did not require any training before usage. As grammar checker we chose Microsoft® Word 2010 (MS Word) since this is a widely used word processor. If there were more than one possible grammatical correction for a candidate translation the first was always chosen. In the case a sentence was high-lighted as extra-grammatical but there were no suggestions on how to correct the translation it was left unchanged. We used the same 2 000 sentences for evaluation as Stymne and Ahrenberg. The BLEU score was calculated by using iBLEU (Madnani, 2011). In our evaluation we went one step further than the replicated study by asking a human translator to analyse the candidate translations as well as the suggested grammatical corrections.

3. Results

3.1 BLEU Scores

In Table 1 the BLEU scores from the original study are given together with the scores from our replication. The last two rows show the BLEU scores for the subset of candidate translations that were corrected by a grammar checker. In both cases Google Translate outperforms the results from Stymne and Ahrenberg with a 35%-increase in BLEU-score. The lower BLEU-scores for the corrected translations might be explained by the fact that these sets only contain a 100 sentences each while BLEU is more reliable for larger evaluation sets (Owczarzak et al., 2007).

3.2 Manual Inspection

The first 100 candidate translations that had a possible correction according to MS Word were manually inspected by the authors and graded. The three possible grades were 'Good', 'Neutral' and 'Bad', the same as in the replicated study. An example of how each grade was used to evaluate the corrections is given below.

MS Word made the suggestion to change the definite noun *utmaningen*, meaning *the challenge*, to the indefinite form *utmaning* in: *Kommissionens utmaningen blir att övertyga parlamentet att den kan skapa dessa garantier*. The suggested change was graded as 'Good' since the *-s* in *Kommissionens*, meaning *the commission's*, marks genitive and for Swedish the rule is that any subsequent noun or noun phrase should use the indefinite form.

An example of a correction that is 'Bad' is given in *Istället är det mer logiskt om varje pigfarmer sätter upp sin egen reserv, d.v.s om alla pigfarmer har sin egen spargris.* Google Translate does not recognise the word *pigfarmer* and transfers it as it is into the Swedish translation. In Swedish *farm* is an English loanword with the plural ending *-er*. MS Word identifies that *pigfarmer* is a compound noun in plural with *farm* as its head. But *varje*, meaning *each*, should be followed by a noun in singular so the grammar checker suggests a correction from the plural form *pigfarmer* into the singular *pigfarm*. Instead of correcting agreement MS Word changes the meaning of the sentence.

In the following translation the underlined *vill*, present and indicative form of *want*, is superfluous; *Och jag vill*

SMT	BLEU	Gram. Check.		
		System	BLEU	Change
Moses	22.18	Granska	22.34	0.16 (0.7%)
Google	29.95	MS Word	29.99	0.04 (0.1%)
Moses	19.44	Granska	20.12	0.68 (3.5%)
Google	23.90	MS Word	24.28	0.38 (1.6%)

Table 1: The BLEU scores for the different systems.

SMT	Gram. check.	Good	Neutral	Bad
Moses	Granska	73	8	19
Google	MS Word	76	3	21

Table 2: The outcome of the manual evaluation of the proposed grammar corrections.

än en gång vill uppriktigt tacka mina kollegor i utskottet för deras samarbete. The suggestion made by MS Word is to replace *vill* with the infinitive form *vilja*. Since changing the word form neither improves nor worsens the fluency the correction was labeled as 'Neutral'.

Just as the above examples suggest, the grammar corrections concerned agreement between adjacent words. In Table 2 the evaluation is presented together with the figures reported by Stymne and Ahrenberg. Since the manual inspections are conducted by different authors the figures are not comparable.

3.3 Analysis by Human Translator

We asked a professional translator between English and Swedish to analyse the translations and the grammatical corrections: *When evaluating the performance of the translator or the grammar checker it is easy to miss the bigger picture. Preserving the intentions of the source text is more than agreement between subject and verb.*

In fact, small improvements as agreement do not make up for the increase in human effort needed to ensure that the grammar checker does not get it wrong. The grammar checker adds a new layer of uncertainty on top of the machine translator's approximation of a translation. The result is that we no longer know where problematic sentences arose. They could be the result of a poor translation by the machine translator, the grammar checker getting it wrong or a combination of the both. Look at the 'Bad' example above. The correction made by the grammar checker hides that pigfarmer *was unknown to the machine translator. A human translator working on the post-processed text could easily miss the mis-interpretation.*

Most importantly, you should never guess! If you are in doubt on how to translate a text you should always get in touch with the customer. Getting it wrong means both a loss of customers and reputation.

4. Discussion

It does not seem that the impact of post-processing the candidate translations with a grammar checker is captured by the BLEU-metrics. Three out of four suggested changes improve the fluency of the translations but for these sentences the increase in BLEU is in our case less than 2%. Our interpretation is supported by the results of Stymne and Ahrenberg as well as by a similar study done by Huet et al. (2010). The latter had an increase from 27.5 on the BLEU-scale to 28.0 after applying a sequence of different post-processing techniques, among them grammar correction.

5. Acknowledgements

The authors would like to thank Carl-Magnus Olsson, Malmö University, and Peter Ljunglöf, University of Gothenburg.

6. References

Johan Carlberger, Rickard Domeij, Viggo Kann, and Ola Knutsson. 2004. The development and performance of a grammar checker for Swedish : A language engineering perspective. *Natural Language Engineering*, 1(1).

Stéphane Huet, Julien Bourdaillet, Alexandre Patry, and Philippe Langlais. 2010. The RALI Machine Translation System for WMT2010. In *ACL 2010 Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pages 109–115, Sweden, Uppsala.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand. Asia-Pacific Association for Machine Translation.

Nitin Madnani. 2011. iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 213 –214, September.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, Rochester, New York, April. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Stymne and Lars Ahrenberg. 2010. Using a Grammar Checker for Evaluation and Postprocessing of Statistical Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, May.

High-Quality Translation: MOLTO Tools and Applications

Olga Caprotti, Aarne Ranta, Krasimir Angelov, Ramona Enache,
John Camilleri, Dana Dannélls, Grégoire Détrez, Thomas Hallgren, K.V.S. Prasad, and Shafqat Virk

University of Gothenburg

Abstract

MOLTO (Multilingual On Line Translation, FP7-ICT-247914, `www.molto-project.eu`) is a European project focusing on translation on the web. MOLTO targets translation that has production quality, that is, usable for quick and reliable dissemination of information. MOLTO's main focus is to increase the productivity of such translation systems, building on the technology of GF (Grammatical Framework) and its Resource Grammar Library. But MOLTO also develops hybrid methods which increase the quality of Statistical Machine Translation (SMT) by adding linguistic information, or bootstrap grammatical models from statistical models. This paper gives a brief overview of MOLTO's latest achievements, many of which are more thoroughly described in separate papers and available as web-based demos and as open-source software.

1. Introduction

MOLTO's mission is to enable **producer-oriented** translation (publishing quality), which means that providers of information can rely on automatically produced translations so much that they can publish them. This is contrasted to **consumer-oriented** translation (browsing quality), which consumers apply to get a rough idea of the content of the original document. On the worldwide web, the consumer scenario is much more common in current machine translation, exemplified by tools such as Google translate and Microsoft Bing.

To achieve publishing quality, MOLTO uses **restricted language**. While consumer oriented translation tools have to cope with any input document, producer-oriented translation tools can assume to only receive documents written in a limited fragment of a language. For instance, e-commerce sites can be restricted to translating product descriptions, order forms, and customer agreements. MOLTO's mission is to make this not only possible but feasible for different users and scenarios, and scalable from small fragments of few languages to larger fragments of large numbers of simultaneous languages.

The diagram in Figure 1. illustrates the goals of MOLTO compared with the goals of consumer-oriented tools. The upper right corner is the ideal of full-coverage full-precision translation, which no current system is even close to achieving. Google translate has to add precision to approach the ideal; full coverage is assumed from the beginning. On the other hand, MOLTO has to broaden coverage; full precision is assumed from the beginning.

The logarithmic scale on the x-axis indicates the number of ”concepts” that are covered. Concepts are the smallest units of translation, such as words, multi-word terms, templates, and constructions. We don't need a precise technical definition of a ”concept” to give an idea of the orders of magnitude. Thus MOLTO is scaling up its translation technology from hundreds to thousands of concepts, but it will still be two or three orders of magnitude behind full-coverage consumer tools.

2. MOLTO's Technology

The main tools under development by the project are:

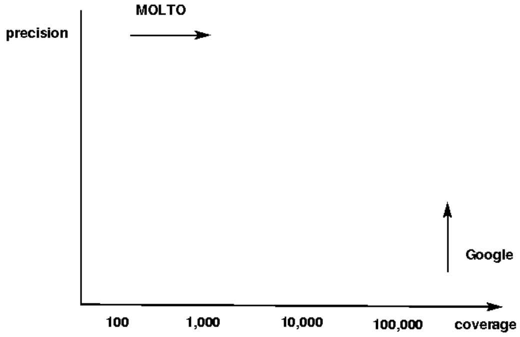


Figure 1: Orthogonal ways to approximate precision and coverage.

The GF grammar compiler (Ranta 2011, Angelov 2012) both a batch compiler for preparing end-user products and an interactive shell for testing grammars during development.

GF Integrated Development Environments comprising a GF-Eclipse plug-in for desktop use (Camilleri 2012) and the GF Simple Editor, a cloud-based editor for GF grammars.

The GF Resource Grammar Library (RGL) providing linguisting functionalities for the morphology, syntax, and lexicon of 25 languages (Ranta 2011). Tools supporting the use of the RGL library include the API synopsis, the source code browser, and the RGL application expression editor.

GF web application interfaces including a small-scale interactive translator that is grammar-driven and a large-scale translator with post-editing support, both implemented using a JavaScript library that enables customization.

Hybrid translation systems using GF starting with a soft-integration GF-SMT architecture (Enache et al. 2012) and more recently with a wide-coverage translator based on resource grammars, robust parsing, and statistical disambiguation (Angelov 2012)

GF tools developed under the MOLTO project are available from the project website and from `http://grammaticalframework.org` under open source licenses (LGPL and BSD).

3. MOLTO's Case Studies

The project has planned a number of application domain in which to test the technologies and the tools developed during its lifetime. Each case study highlights key features of the approach of MOLTO to multilingual translation.

The MOLTO Phrasebook (Détrez et al. 2012) is the controlled natural language of touristic phrases. It covers greetings and travel phrases such as "how far is the airport from the hotel" in 18 languages. The translations show the kind of quality that can be hoped for when using a GF grammar that can handle disambiguation in conveying gender and politeness, for instance from English to Italian. It is available both on the web and as a stand-alone, offiline Android application, the PhraseDroid.

A different kind of controlled natural language is mathematical text, tackled in the Mathematical Grammar Library (Saludes and Xambó 2011). Examples are educational courseware in mathematics, especially exercises but also command languages for computational software systems such as Sage. In the GFSage software application, a command-line tool is takes commands in natural language, has them executed by Sage, and returns the answers in natural language. Combined with speech recognition and synthesis, this application demonstrates how a MOLTO library can add multimodality to a system originally developed with keyboard input/output as user interface.

To demonstrate the MOLTO Knowledge Reasoning Infrastructure, the Patent retrieval prototype shows examples of queries in natural language to a collection of patents in the pharmaceutical domain. Users can ask questions like "what are the active ingredients of AMPICILLIN" in English, French, and German. At present the online interface allows to browse the retrieved patents and returns the semantic annotations that explain why any particular patent has matched the user's criteria. Similar technology for knowledge retrieval is being applied also in the case of cultural heritage, namely with descriptions of artefacts from the museum of Gothenburg, in order to allow multilingual query and retrieval in 6 languages (Dannélls 2012). For this task, an ontology has been created together with a GF application grammar.

In 2011 the MOLTO consortium was extended with two more partners with two new scenarios. The first aims to demonstrate how a collaborative editing environment can be kept synchronous and multiligual by a generalization of the AceWiki software platform. AceWiki has already been extended to allow multilingual constrained natural language input and editing, driven by multilingual GF grammars (Kaljurand 2012). The second addresses the challenges of creating consensus-based, multi-authored ontologies that capture business processes by linguistically untrained stakeholders across business disciplines. GF is used to create multilingual grammars that enable transparent multilingual verbalisation (Davis et al. 2012).

4. Some Conclusions

MOLTO's tools are becoming a standard approach in controlled language implementation and ontology verbalization, which is shown for instance by the proceedings of the last two Controlled Natural Language workshops (Fuchs and Rosner 2012, Kuhn and Fuchs 2012). The potential of this field is shown in particular by the growing interest in multilingual semantics web.

Scaling up to a larger coverage is show by the hybrid system of Enache and al. (2012), with expected loss of quality but still an improvement over an SMT baseline. This is another growing field, where in particular GF's ability to cope with rare language pairs (with insufficient statistical data) can be exploited.

Acknowledgements. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n:o FP7-ICT-247914.

5. References

K. Angelov. *The Mechanics of the Grammatical Framework.* PhD Thesis, Chalmers University of Technology, 2012.

K. Angelov and R. Enache. Typeful Ontologies with Direct Multilingual Verbalization. In Fuchs and Rosner (2012).

J. Camilleri. An IDE for the Grammatical Framework, *Proceedings of FreeRBMT12*, to appear, 2012.

G. Détrez, R. Enache, and A. Ranta. Controlled Language for Everyday Use: the MOLTO Phrasebook. In Fuchs and Rosner (2012).

D. Dannélls. On generating coherent multilingual descriptions of museum objects from SemanticWeb ontologies, *International Conference on Natural Language Generation (INLG)*, 2012.

D. Dannélls, R. Enache, M. Damova, and M. Chechev. Multilingual Online Generation from Semantic Web Ontologies. *WWW2012, 04/2012, Lyon, France*, (2012).

B. Davis, R. Enache, J. van Grondelle and L. Pretorius. Multilingual Verbalisation of Modular Ontologies using GF and lemon.

R. Enache, C. España-Bonet Cristina, A. Ranta, and L. Màrquez , A Hybrid System for Patent Translation, *EAMT 2012*.

N. Fuchs and M. Rosner (eds), *CNL 2010 Proceedings*, Springer LNCS/LNAI vol. 7175, 2012.

K. Kaljurand, General Architecture of a Controlled Natural Language Based Multilingual Semantic Wiki. In Kuhn and Fuchs (2012).

T. Kuhn and N. Fuchs (eds), *CNL 2012 Proceedings*, Springer LNCS/LNAI vol. 7427.

A. Ranta. *Grammatical Framework: Programming with Multilingual Grammars*, CSLI Publications, Stanford, 2011.

J. Saludes, and S. Xambó, *The GF Mathematics Library*, Proceedings First Workshop on CTP Components for Educational Software (THedu'11), Electronic Proceedings in Theoretical Computer Science, Number 79, Wrocław, Poland, p.102–110, (2011) .

Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care

Hercules Dalianis, Martin Hassel, Aron Henriksson, Maria Skeppstedt
Department of Computer and Systems Sciences (DSV), Stockholm University
Forum 100, 164 40 Kista, Sweden
{hercules, xmartin, aronhen, mariask}@dsv.su.se

Abstract

The care of patients is well documented in health records. Despite being a valuable source of information that could be mined by computers and used to improve health care, health records are not readily available for research. Moreover, the narrative parts of the records are noisy and need to be interpreted by domain experts. In this abstract we describe our experiences of gaining access to a database of electronic health records for research. We also highlight some important issues in this domain and describe a number of possible applications, including comorbidity networks, detection of hospital-acquired infections and adverse drug reactions, as well as diagnosis coding support.

1. Introduction

Health care costs increase rapidly with aging populations. However, at the same time, information technology is making health care more efficient. Information technology has made a significant contribution to improving health care, e.g. in the form of digitized radiology images, monitoring equipment for heart and blood pressure and automatic chemical analysis of body fluids. Moreover, health records have become digitized in many parts of the world, which also allows patient data to be stored in centralized databases. These patient databases contain documentation written by experienced health care personnel and record a patient's symptoms, diagnoses, treatments, prescribed drugs, etc. In countries where each patient has a unique social security number, as in Scandinavia, longitudinal studies of patients can be undertaken across clinics. In short, clinical documentation contains valuable information concerning the treatment of real patients; however, this is rarely available for research due to its sensitive nature.

2. Our Data

2.1 Obtaining Data

In order to gain access to clinical data for research, ethical permission from the local vetting board is required. The ethical application must contain a clear research question and also describe how the data will be treated and stored. Furthermore, it must be shown to the hospital management that the research will be beneficial for them and that their data will not be misused.

2.2 Storing Data Securely

In our case, the data were obtained without patient names and the social security numbers were replaced with unique serial numbers. However, the textual fields in the data may still contain sensitive identifying information, e.g. names, phone numbers and addresses. Our data are stored on encrypted servers without network access in a locked and alarm-equipped server room.

2.3 Experts in Medicine

Medical domain experts are needed to interpret the noisy text due to the professional language that pervades health records. We have had access to three senior physicians for the annotation and interpretation of our data.

2.4 Features of the Stockholm EPR Corpus

The Stockholm EPR (Electronic Patient Record) Corpus contains data from over 512 clinical units from Stockholm City Council encompassing the years 2006–2010 and over one million patients. The whole corpus contains over 500 million tokens. Certain statistics for the years 2006–2008 can be found in Dalianis et al. (2009).

3. Other Clinical Databases

The following is a brief description of five different clinical databases from three countries that have been made available for research:

- The i2b2[1]corpus is a clinical corpus consisting of approximately 1,000 notes in English.
- The CMC[2] clinical corpora consists of 2,216 records in English from children's radiology departments, which have been automatically and manually de-identified (Pestian et al., 2007).
- The De-id[3] corpus consists of 412,509 nursing notes and 1,934 discharge summaries written in English.
- A Finnish clinical corpus [4] containing 2,800 sentences (17,000 tokens) from nursing notes, which have been manually anonymized.
- THIN[5] contains the records of almost 11 million patients from general practices. The records are from the years 1986–2003 and are written in English. The database specializes in pharmacoepidemiology (see Lewis et al., 2007).

4. Applications

There is a nice overview of NLP applications using health record texts in Meystre et al. (2009). Below are some examples of what our research group has worked on using our clinical database, the Stockholm EPR Corpus.

[1] http://www.i2b2.org
[2] http://computationalmedicine.org/catalog
[3] http://www.physionet.org/physiotools/deid
[4] http://bionlp.utu.fi/clinicalcorpus.html
[5] http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database

4.1 Comorbidity Networks

Medical records databases can be used to create comorbidity networks, i.e. a network of patients suffering from at least two diseases. These can be explored from the perspective of age, gender and specific ICD-10 diagnosis codes; an example can be seen in our demo version of Comorbidity View[6], which uses structured data from the Stockholm EPR Corpus (2006–2008). This approach can also be used to align diseases, drugs and patient groups. By extracting relevant information from the free text, this part of the health record can be used to build various co-occurrence networks. We have annotated clinical text and developed tools based on machine learning techniques that recognize symptoms, diagnoses and drugs in the free text (Skeppstedt et al., 2012). These tools are also able to recognize if the mentioned symptoms and diagnoses are in a negated or speculative context (Velupillai, 2012).

4.2 Detecting Hospital Acquired Infections

Around ten per cent of all in-patients are affected by hospital acquired infections (Humphreys & Smyths, 2006), however many of these hospital-acquired infections (HAI) are not reported correctly to the hospital management. One way to achieve automatic detection of HAIs is to have computers analyse the free text and structured information in the patient records and search for indications of HAIs, e.g. that a patient has been treated with a particular antibiotic drug after the use of a catheter (Proux et al., 2011). We are currently annotating 180 Swedish patient records, some of which contain HAIs, in order to evaluate our rule-based and machine learning tools that retrieve information from the records and determine whether they contain a possible HAI or not.

4.3. Detecting Adverse Drug Reactions

Pharmaceutical drugs are developed through laboratory tests and clinical trials; however, these methods are very expensive. To reduce costs one is also doing simulation using mathematical models but such methods are not always sufficiently robust to identify the effects and side effects of drugs. We plan to use our medical database to confirm the known effects of drugs, but also to find unknown adverse drug reactions (ADR). A initial approach is presented in Henriksson et al. (2012), where Random Indexing (a form of distributional lexical semantics) is applied to a part of the Stockholm EPR Corpus in order to extract drug-symptom pairs. Around 50 per cent of the drug-related words were conceivable ADRs, while ten per cent were known and documented ADRs.

4.4. Diagnosis Coding Support

Accurate coding of diagnoses enables statistical monitoring of symptoms, diseases and treatments. This is, however, a time-consuming and thus costly administrative task that has led to attempts to provide computer-aided coding support. We are working on a method that uses Random Index-based word spaces, containing co-occurrence information of textual units and diagnosis codes, to recommend possible codes for an uncoded clinical note (see, e.g., Henriksson et al., 2011).

[6] http://dsv.su.se/en/research/health/comorbidityview/demo/

5. Conclusions

Obtaining ethical permission and secure access to the data is probably the most difficult part of this research for the researcher. Other challenges involve the problem of interpreting and annotating the clinical data which requires the aid of medical domain experts such as physicians and nurses. The noisy data are also an obstacle for our natural language processing tools. However, given the extra resources afforded by the records, the domain presents interesting and relatively uncharted territory for language technology that can contribute to improved health care.

References

Dalianis, H., Hassel, M. and Velupillai, S. (2009). The Stockholm EPR Corpus – Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249.

Henriksson, A., Hassel, M. and Kvist, M. (2011). Diagnosis Code Assignment Support Using Random Indexing of Patient Records – A Qualitative Feasibility Study. In Proceedings of AIME, 13th Conference on Artificial Intelligence in Medicine, Springer-Verlag, pp. 348–352.

Henriksson, A., Kvist, M., Hassel, M. and Dalianis, H. (2012). Exploration of Adverse Drug Reactions in Semantic Vector Space Models of Clinical Text. In Proceedings of ICML Workshop on Machine Learning for Clinical Data Analysis.

Humphreys, H. and Smyths, E.T.M. (2006). Prevalence surveys of healthcare-associated infections: what do they tell us, if anything? *Clin Microbiol Infect* 12, pp. 2–4.

Lewis, J., Schinnar, R., Warren, B., Bilker, W., Wang, X. and Strom, B. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidem Drug Safe*, pp. 393–401.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C. and Hurdle, J.F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform,* pp. 128–144.

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. and Duch, W. (2007). A shared task involving multi-label classification of clinical free text, BioNLP: Biological, translational, and clinical language processing, ACL, pp. 113–120.

Proux, D., Hagège, C., Gicquel, Q., Pereira, S., Darmoni, S., Segond, F. and Metzger, M-H. (2011). Architecture and Systems for Monitoring Hospital Acquired Infections inside a Hospital Information Workflow. In Proceedings of the Workshop on Biomedical Natural Language Processing, RANLP-2011, pp. 43–48.

Skeppstedt, M., Dalianis, H., Kvist, M. and Nilsson, G.H. (2012). Detecting drugs, disorders and findings in Swedish Clinical text using CRF, (forthcoming).

Velupillai, S. (2012). Shades of Certainty: Annotation and Classification of Swedish Medical Records. PhD thesis, Department of Computer and Systems Sciences, (DSV), Stockholm University.

The gravity of meaning: Physics as a metaphor to model semantic changes

Sándor Darányi, Peter Wittek

Swedish School of Library and Information Science
University of Borås
Allegatan 1, Borås S-50190, Sweden
sandor.daranyi@hb.se, peterwittek@acm.org

Abstract

Based on a computed toy example, we offer evidence that by plugging in similarity of word meaning as a force plus a small modification of Newton's 2nd law, one can acquire specific "mass" values for index terms in a Saltonesque dynamic library environment. The model can describe two types of change which affect the semantic composition of document collections: the expansion of a corpus due to its update, and fluctuations of the gravitational potential energy field generated by normative language use as an attractor juxtaposed with actual language use yielding time-dependent term frequencies. By the evolving semantic potential of a vocabulary and concatenating the respective term "mass" values, one can model sentences or longer strings of symbols as vector-valued functions. Since the line integral of such functions is used to express the work of a particle in a gravitational field, the work equivalent of strings can be calculated.

1. Introduction

In support of arguments presented by Baker (Baker, 2008) and Paijmans (Paijmans, 1997), we offer the following thought experiment. As the ongoing debate between prescriptive and descriptive linguistics bears witness to it, language is not used in a vacuum and its proper functioning rests on a now unspoken then codified consensus, maintained by tradition. Therefore while attempting to model and incorporate word or sentence meaning, language use based models of text categorization (TC) or information retrieval (IR) cannot disregard this influence.

In what follows, we propose a new descriptive model which factors in this consensus as one of its components. In a metaphorical fashion, we use key concepts from classical mechanics to model word and sentence meaning as they manifest themselves in an evolving environment, i.e. language change is ab ovo taken into consideration.

Modelling an expanding text corpus by Salton's dynamic library (Salton, 1975), we depart from a vector space able to host both exactly and inexactly (regionally) located semantic content such as term vectors constituting document and query vectors (Gärdenfors, 2000; Erk, 2009). If we turn our attention to a physical metaphor, we may say that the force between the interacting particles (words, sentences, or other units of meaning) decays at least quadratically with distance. This means that their impact is negligible beyond a region, hence the model naturally incorporates semantic regionality.

In accord with the metaphor, energy as work capacity is stored in structures in nature, whereas meaning as the capacity to carry out work on one's mind is likewise stored in ontological and syntactic structures in language. This is underpinned by the fact that the same type of eigen decomposition is used in statistical physics to develop a deep understanding of a system under study, and in text analysis to assign term groups to certain latent variables.

Using concepts like location, distance, time, velocity, acceleration and force, we depart from identifying index term "mass" in an evolving document corpus. To this end, we plug in similarity as a force holding terms with a similar word meaning together, i.e. providing coherence for lexical fields (White, 2002). By a slightly modified version of Newton's 2nd law, specific term "masses" clearly emerge (Tables 1-2), although their empirical status needs more investigations. On the other hand, including the prescriptive-normative part of language use as an Earth-like "mass" in the model by a constant, this and term "masses" with time-dependent term positions yield the semantic analogue of gravitational potential energy, a field. In this field, gradual dislocations of term "masses" represented as position vectors reveal their respective direction vectors as well.

2. Implications

Given the above "semantic gravitational constant", two observations are imminent. First, the changing field index terms and normative consensus jointly generate will be influenced by forces outside of the system only, i.e. by updates of the text corpus such as adding new documents with new vocabularies to it. On the one hand, such updates correspond to the kinetic component of the Hamiltonian of the system which sums up its energy balance. On the other hand, this kinetic component changing the system at t_n will be added to its potential component at t_{n+1} whereby we can identify the Hamiltonian of such systems in general. This continuous net potential gain goes back to the fact that accumulated information as the potential energy part of the Hamiltonian is not reduced by being copied out to carry out work, which is a strong divergence from the rules of nature. Rather, iteratively copied out potential energy seems to underlie a series of information explosions feeding back to the same storage system (i.e. dynamic library, text corpus). This positive feedback leads to cycles of information generating new information which then becomes harvested for the same collection.

Secondly, with word semantics modelled on a physical field, sentences as word sequences can be represented by concatenated term position and direction vectors, i.e. vector valued functions (VVF). As work in physics is the prod-

$t = 0$	Doping	Football	Performance	Skiing	Training
d_1	5	2	0	0	0
d_2	4	0	0	3	1
d_3	0	0	4	0	5
d_4	6	0	2	0	0
d_5	0	3	0	0	4
$t = 1$					
d_1	5	2	0	0	0
d_2	4	0	0	3	1
d_3	0	0	4	0	5
d_4	6	0	2	0	0
d_5	0	3	0	0	4
d_6	2	3	0	1	1
d_7	1	0	0	4	5
$t = 2$					
d_1	5	2	0	0	0
d_2	4	0	0	3	1
d_3	0	0	4	0	5
d_4	6	0	2	0	0
d_5	0	3	0	0	4
d_6	2	3	0	1	1
d_7	1	0	0	4	5
d_8	5	6	1	1	0
d_9	2	1	1	3	0

Table 1: Evolution of an indexing vocabulary over time

	Doping	Football	Performance	Skiing	Training
v_1	9	9	0	25	36
v_2	49	49	4	16	0
a	40	40	4	-9	-36
F	1.56	1.28	1.24	1.35	1.37
m	0.039	0.032	0.31	0.15	0.038

Table 2: Calculation of term mass over t_0-t_2

uct of force and distance, and the work a particle carries out in a physical field is equals the line integral of its trajectory, a VVF, this way we can compute the work equivalent of a sentence or a longer symbolic string. Because sentence representation by tensor products (Aerts and Gabora, 2005), or holographic reduced representations (Plate, 1991; De Vine and Bruza, 2010) is a less intuitive process, sentences as concatenated or convoluted VVFs in Hilbert space is a step toward a simple and comprehensive new frame of thought. This representation has the following advantages:

- It utilizes physics as a metaphor to model the dynamics of language change;
- It demonstrates the connection between sentence structure and work carried out in a field based on classical (Newtonian) mechanics, i.e. is feasible to quantify the work content in documents ;
- It models such a field as the gravitational potential energy terms possess in the presence of language norms, with similarity as a force between pairs of them as the gradient of the above potential;
- It naturally bridges the gap between language analysis and language generation.

3. Future work

The current model yields variable term mass over observation periods which departs from its roots in classical mechanics. Although ultimately language may show different ”symptoms of behaviour” as physics does, we are working on an alternative to yield constant term mass values, leading to scalability tests and evaluation of the new model.

4. References

D. Aerts and L. Gabora. 2005. A theory of concepts and their combinations II: A Hilbert space representation. *Kybernetes*, 34(1-2):192–221.

Adam Baker. 2008. Computational approaches to the study of language change. *Language and Linguistics Compass*, 2(3):289–307.

L. De Vine and P. Bruza. 2010. Semantic oscillations: Encoding context and structure in complex valued holographic vectors. In *Proceedings of QI-10, 4th Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pages 11–13, Arlington, VA, USA, November.

K. Erk. 2009. Representing words as regions in vector space. In *Proceedings of CoNLL-09, 13th Conference on Computational Natural Language Learning*, pages 57–65, Boulder, CO, USA, June.

P. Gärdenfors. 2000. *Conceptual spaces: The geometry of thought*. The MIT Press.

H. Paijmans. 1997. Gravity wells of meaning: detecting information-rich passages in scientific texts. *Journal of Documentation*, 53(5):520–536.

T.A. Plate. 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of IJCAI-91, 12th International Joint Conference on Artificial Intelligence*, pages 30–35, Sydney, Australia, August.

G. Salton. 1975. Dynamic information and library processing.

H.D. White. 2002. Cross-textual cohesion and coherence. In *Proceedings of the Workshop on Discourse Architectures: The Design and Analysis of Computer-Mediated Conversation*, Minneapolis, MN, USA, April.

Coordinating spatial perspective in discourse

Simon Dobnik

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 200, 405 30 Göteborg
simon.dobnik@gu.se

1. Introduction

Understanding and generating spatial descriptions such as “to the left of” and “above” is crucial for any situated conversational agent such as a robot used in rescue operations. The semantics of spatial descriptions are complex and involve (i) perceptual knowledge obtained from scene geometry (Regier and Carlson, 2001), (ii) world knowledge about the objects involved (Coventry et al., 2001), and (iii) shared knowledge that is established as the common ground in discourse. Dialogue partners coordinate all three types of meaning when describing and interpreting visual scenes.

One example of (iii) is the assignment of perspective or the reference frame (RF). For example, the table may be “to the left of the chair”, “to the right of the chair”, “behind the chair” or “South of the chair”. The RF, may be described linguistically “from your view” or “from there” but in a spontaneous conversation it is frequently omitted. Instead, it is integrated in the content of the conversation as a discourse variable which is applied over several turns and even over several speakers. The RF may also be inferred from the perceptual context if given some configuration of the scene a spatial description is true only in that RF. It follows that when interpreting and generating spatial descriptions humans rely on verification of spatial templates in different RFs which requires considerable computational complexity (Steels and Loetzsch, 2009).

The perspective is grounded by some point in the scene called the *viewpoint* (VPT). There are three ways in which the VPT is set in human languages (Levinson, 2003): (i) *relative RF:* by some third object distinct from the located and reference objects (the speaker, the hearer, the sofa); (ii) *intrinsic RF:* by the reference object itself (the chair); or (iii) *extrinsic RF:* by some global reference point (the North). Sometimes (mostly for route descriptions) a distinction is made between speaker-oriented (egocentric) and external (allocentric) perspective or between route and survey perspective but this is a less specific distinction. The geometric spatial templates are projected within the framework defined by the VPT (Maillat, 2003).

2. Reference frames in conversation

Watson et al. (2004) show experimentally that (i) participants are significantly more likely to use an intrinsic RF after their partner used an intrinsic RF, compared when the partner used a relative RF (with the speaker as the VPT); (ii) participants are significantly more likely to use intrinsic RF when the objects are aligned horizontally (their typical alignment in the world) than when they are aligned vertically; (iii) the alignment of the RFs is not due to the lexical priming caused by using the same preposition. Andonova (2010) shows for the map task that overall partners align with the primed route or survey perspective set by the confederate if priming is consistent – when the confederate changes the perspective only once in the middle of the session. On the other hand, if the confederate regularly alternates between the perspectives their partner has nothing to prime to. The self-assessed spatial ability (using a standardised test) is also important – low ability participants only align with the primed perspective when the switch is from the survey to the route perspective which is otherwise also the most frequently used one.

3. Towards a more natural spatial dialogue

Our interest is to implement these and similar strategies as information state update rules in a dialogue manager such as GoDiS (Larsson, 2002). In such a model each conversational agent must keep a record of their own RF and that of their partner in the common ground. The RFs are updated following perceptual verification and an alignment strategy. The proposal is a move towards a more natural interpretation and generation of projective spatial descriptions in an artificial conversational agent compared to our previous attempt where the RF parameters were not specifically included in the model but some RF knowledge has nonetheless been learned with machine learning. We proceed as follows:

1. Collect a corpus of dialogue interactions containing projective spatial descriptions made in a room scene.
2. Annotate the dialogue utterances with an XML annotation scheme which identifies perceptual states, objects in focus, utterances, turns, speakers, located objects, RFs, VPTs, spatial relations, ref. objects, etc.
3. Replicate the literature findings on the RF usage in our dataset.
4. Repeat the experiments from (1) but where one of the participants is a dialogue manager following an RF strategy. Allow humans conversational partners to rate the performance of the system.
   (a) Always use the relative RF to yourself.
   (b) Always align to the RF used by your partner in the previous turn.
   (c) For each turn select the RF randomly.
   (d) Keep a randomly chosen RF for n turns, then change.

   To prevent over-agreement with the system the evaluators should, ideally, compare pairs of strategies and select the preferred one.

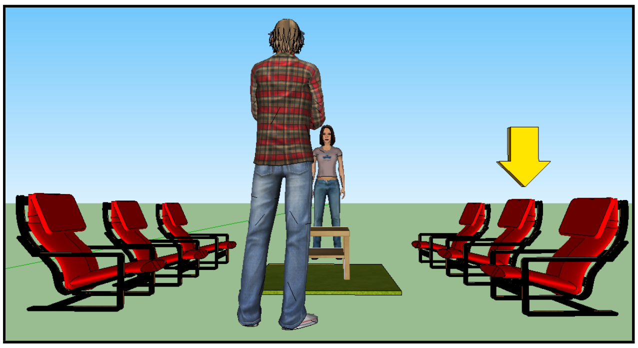


Figure 1: A scene one participant sees during a conversation in the 2nd pilot study. The arrow indicates the object the location of which this participant should describe.

We collect our data and later test the interaction in an online experimental environment specifically developed for this purpose (`http://goo.gl/8KLja`). Participants may create sessions to which they invite other participants and complete them interactively in their own time. During a session each participant sees a 3d generated image of a room containing some furniture. The image also contains two avatars: the one with their back towards the participant is the participant and the one facing the participant from the opposite side of the room is their partner (Figure 1). This is explained to the participants in the instructions and different representations are used to avoid the confusion. The other participant sees the room from the opposite side. The participants communicate via a text chat interface which allows unrestricted entry of text and also logs and partially annotates both the conversation and the perceptual information in the background.

4. Results

By the time of writing this abstract we conducted two pilot studies for which we completed stages 1 to 3 of our plan. In the first pilot study (7 conversations in Slovenian) we used a room with four distinct entities (two participants, a chair and a sofa) arranged around a table in the middle which was placed on a carpet. We instructed the participants to talk about the location of the objects in the scene. Although this method was good in encouraging spontaneous conversations it had two shortcomings: (i) the participants produced less spatial descriptions than desired (11.9 per conversation) as they also discussed their opinions about the objects, etc.; and (ii) they spontaneously took on roles where one was asking questions and the other was giving answers and therefore the conversations included were very few cases of interaction that we were looking for. To overcome the difficulties from the first study we designed a second pilot study (10 conversations in Slovenian) for which we (i) only used one kind of objects (the chairs), (ii) restricted the conversational interaction to pair of turns where in the first turn one participant describes which chair they chose (one is automatically selected for them and marked with an arrow as shown in Figure 1) and then in the second turn their partner selects that chair on their view of the room. The roles are reversed in the next turn. Thus, we get a series of dialogue turns from which we record (i) speaker's strategy for RF choice; (ii) the hearer's understanding of the description. The latter is important as a particular description may be true under more than one RF.

A manual analysis of the data obtained so far confirms that participants align their perspective but only if one participant uses a particular perspective consistently over more than one turn, then the other would follow (priming). Our explanation is that the second speaker assumes that a particular perspective is important in the conversation and should therefore be made part of the common ground. Further we observe that speakers not only align perspectives but also the way the scene is described syntactically. While in the first trials participants may frequently omit the explicit description of perspective and align to the perspective of the other, the structured environment of the second trials forces them to use definitions such as “from your side” nearly all the time even if they are aligned. They may also omit the explicit definition and align to the fact that each participant is describing from its own perspective.

5. Further work

In the time leading to the conference we hope to continue to collect conversations online (in Slovenian, English and Swedish), tag them and integrate them in an automatic agent that will be used for step 4. Note that the method allows us to collect a set of best referring expressions for each object together with all their semantic properties which means that the descriptions can be conveniently applied in generation.

6. References

Elena Andonova. 2010. Aligning spatial perspective in route descriptions. In Christoph Hölscher, Thomas Shipley, Marta Olivetti Belardinelli, John Bateman, and Nora Newcombe, editors, *Spatial Cognition VII*, volume 6222 of *Lecture Notes in Computer Science*, pages 125–138. Springer Berlin, Heidelberg.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of memory and language*, 44(3):376–398.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, University of Gothenburg.

Stephen C Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*, volume 5. Cambridge University Press, Cambridge.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, Committee for Comparative Philology and General Linguistics, University of Oxford, Oxford, UK, May.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

Matthew E. Watson, Martin J. Pickering, and Holly P. Branigan. 2004. Alignment of reference frames in dialogue. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, USA, August.

Are prepositions and conjunctions necessary in health web searches?

Ann-Marie Eklund

Department of Swedish, University of Gothenburg
Gothenburg, Sweden
ann-marie.eklund@gu.se

Abstract

People searching for information via public health web sites may differ in medical knowledge and internet experience. Hence, they may show different search behaviour with implications for both search engine optimisation and log analysis. In this paper we address the question of how prepositions and conjunctions are used in search, concluding that they do occur in different contexts, but do not impact a semantic mapping for query optimisation and analysis.

1. Introduction

Today, there is an increasing interest in using the internet as a means for health care providers to interact with the general public. These interactions often take place via web portals providing tools to administrate personal health care activities and provide information on diseases, symptoms and treatments.

In a series of papers (Eklund, 2012; Oelke et al., 2012), we have studied the use of the Swedish health web site <http://hitta.vgregion.se> (now part of the site 1177.se), which is a major source of health information for 1.5 million inhabitants of the Swedish region Västra Götaland.

Our interest in studying the use of this web site is twofold. Firstly, since searching health portals like 1177.se is often a major way for the public to access relevant information, these portals need to be optimised to consider the search behaviour and used language of its users, who may not have Google or medical experience. Secondly, as has been presented by for instance Hulth et al. (2009), the use of health portals may be related to media coverage and the current health status in society. Hence, to analyse this use it is important that the data is semantically normalised to take into account aspects like different ways of expressing the same concept, for instance, the popular use of the term "svininfluensa" ("swine flu") in comparison to the official name "the new influensa", or the use of the term vinterkräksjuka ("winter vomiting disease") as an expression for stomach flu in general.

In this paper we will study the use of prepositions and conjunctions in search queries. The latter category is interesting to study since it is relevant to know if conjunctions are used in a semantically important way or just to indicate logical connectives as for instance "and". Similarly, understanding the use of prepositions may also provide insights into their roles and potential importance in terminology based search optimisation.

2. Material and method

We have studied a search log from the Swedish health web site <http://hitta.vgregion.se> (now part of the site 1177.se) covering the time period from June 2010 to September 2011. The site is the official health care portal of the region Västra Götaland in western Sweden and is a major source of health information for the region's 1.5 million inhabitants. The log contains 230,754 interactions, that is, pairs of queries and chosen results, grouped into sessions. It contains 31,054 queries, most of which consist of one to three words. For a more detailed description of the search log, see (Eklund, 2012; Oelke et al., 2012).

Since our interest is improved understanding of search behaviour, we utilise a method based on mapping query terms or phrases to their semantic counterparts, thereby providing a normalised semantic based corpus for further analysis.

In the UMLS (Unified Medical Language System) Metathesaurus (www.nlm.nih.gov/research/umls), terms from medical and health related vocabularies are linked to concepts. Each concept is linked to at least one semantic type. We mapped the queries to concepts and to the semantic types of these concepts. We tried to map the whole query to a concept, even when the query consisted of more than one word. E.g. the whole query "propp i lungan" ("pulmonary embolism") could be mapped to a UMLS concept and to its semantic type Disease or Syndrome. If no concept was found, the query was split into words and we tried to find a matching concept for each word. The mapping was done using the methods described in (Eklund, 2012; Kokkinakis and Gerdin, 2009).

3. Results

3.1 Prepositions

The most common preposition in the search log is "i" ("in") and it is used in 905 queries. Many of these (231 queries) contain "ont i" ("pain in") and a specification of where, e.g. "ont i magen" ("stomach ache"), "ont i halsen" ("sore throat"). In nearly all cases "ont" ("pain") is expressed with the preposition, e.g. "ont i magen" ("pain in the stomach"). In very few cases the preposition has been omitted, e.g. "ont magen" ("pain stomach") or "ont axeln" ("pain shoulder"). The preposition "i" is often followed by the semantic type Body Part, Organ, or Organ Component. Hence, it is used to relate symptoms and body parts, e.g. "ont i njure" ("pain in kidney"). For some queries with "i" it was possible to map the whole query to a concept and its semantic type, e.g "propp i lungan" ("pulmonary embolism") and "ont i magen" were mapped to the semantic type Sign or Symptom. However, worth noticing is that not only for instance “ont i magen”, but also “ont magen” could be mapped to the semantic type Sign or Symptom, i.e. using or omitting “i” in this case made no difference. There are 50 queries containing an expression of “i” followed by a geographic area, e.g. “sjukhuset i skövde” (“the hospital in skövde”) or “rehabilitering i västra götaland” (“rehabilitation in västra götaland”), but it is much more common (about 2,500 queries) to express a geographic location without using “i”, e.g. “bvc lidköping” (“child health care lidköping”) or “psykiatri mölndal” (“psychiatry mölndal”).

Following “i” the most commonly used preposition is “på” (“on”/“at”) (171 queries). The use of this preposition is mainly to connect findings, symptoms or abnormalities with diseases, or to emphasise their occurrence at a given part of body or organ. In the latter case, the whole phrase containing the preposition may be seen as a symptom, e.g. “blåsor på tungan” (“blisters on the tongue”). However, eventhough many of these phrases may be considered to have a given semantics, mostly they express a semantic relation and not a concept. Examples where this mapping may be inaccurate are “fickor på tarmen” (“pockets on the intestines”) and “brist på b12” (“lack of b12”), which should be mapped to “diverticulitis” and “vitamin b12 deficiency” respectively.

The third most used preposition is “av” (“of”) (144 queries). This preposition is often used in the context of treatment of diseases, e.g. “behandling av hepatit c” (“treatment of hepatitis c”), and parts of organs “vänstra delen av buken” (“the left part of the abdomen”). For these two types of expressions it is difficult to establish semantic mappings. In the first case it expresses a semantic relation. In the second example we have meronymy, which in terminologies is often not expressed in terms of “left” and “right”, but of their medical terms.

3.2 Conjunctions

Two conjunctions of importance considering semantic mappings and search engine optimisations are “och” (“and”) and “eller” (“or”). In the analysed 1177.se search log the conjunction “och” has been used in 254 queries and “eller” only in 14 queries. The overwhelming part of the use of “och” is as a logical connective, e.g. the information seeker expresses an interest in both “nausea” and “backpain” (“illamående och ryggont”). Another way of using “och” is to express e.g. pain in the back and pain in the stomach as “ont i ryggen och magen” (“pain in the back and in the stomach”). For this type of syntactic pattern it is more challenging to achieve an accurate semantic mapping.

4. Discussion

Our initial analysis shows that both prepositions and conjunctions are used when querying the Swedish public health site 1177.se. Considering the most common preposition “i”, it is often used to relate terms of different semantic types, e.g. signs or symptoms and body parts, hence to manifest the origin of a symptom. In other words, the preposition may be seen as a marker of a semantic relation. Eventhough the terms in a terminology generally do not describe relations, in cases like this the relation has become a concept in itself in the UMLS (SNOMED terminology), thereby reducing the importance of the existence of the preposition when mapping from the query terms to their semantic counterparts.

In comparison to the use of the preposition “i” when considering symptoms, in connection with geographic areas the information seeker almost never find it necessary to express that a health care facility is located “in” a particular city or area. One explanation for this difference may be that in the case of symptoms e.g. “ont i” (“pain in”) is an established expression.

Considering the less commonly used prepositions “på” and “av”, the first one is used in a similar way as “i”, relating disease information and parts of the body. However, as in the case of “i”, there are also instances where the complete phrase including “på” should be mapped to a single concept. Often this type of phrases could be rephrased as a compound and thereby be mapped to a concept. In the case of “av”, it often relates treatments to parts of the body or expresses meronymy not expressed by medical terminology, thereby being more difficult to map to accurate concepts.

When conjunctions are used, it is almost always to express logical connectives, thereby not leading to any mapping problems. However, we have also seen examples of specific syntactic patterns which can be a problem for semantic mappings.

To return to the topic of this paper, if prepositions and conjunctions are necessary in health web searches, our analysis indicates that - eventhough there are a few examples like specific syntactic patterns, meronymy and prepositional phrases used instead of a compound - prepositions and conjunctions may be treated as stopwords and excluded from semantic analysis. Hence, they do not need to be taken into specific consideration in the context of health web searches.

5. Acknowledgments

The research presented here was supported by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken. The author would also like to thank Findwise (findwise.com) for providing the search log.

6. References

Ann-Marie Eklund. 2012. Tracking changes in search behaviour at a health web site. In *Proceedings of the 24th European Medical Informatics Conference*.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web queries as a source for syndromic surveillance. *PLoS One*, 4(2):e4378.

Dimitrios Kokkinakis and Ulla Gerdin. 2009. Issues on quality assessment of SNOMED CT subsets - term validation and term extraction. In *Workshop: Biomedical Information Extraction (Recent Advances in Natural Language Processing)*.

Daniela Oelke, Ann-Marie Eklund, Svetoslav Marinov, and Dimitrios Kokkinakis. 2012. Visual analytics and the language of web query logs - a terminology perspective. In *The 15th EURALEX International Congress (European Association of Lexicography)*.

Ontology matching: from PropBank to DBpedia

Peter Exner **Pierre Nugues**

Department of Computer Science, Lund University, Sweden
peter.exner@cs.lth.se, pierre.nugue@cs.lth.se

1. Introduction

In ontology matching, disparate ontologies expressing similar concepts are aligned, enabling tasks such as data and ontology integration, query answering, data translation, etc. (Pavel and Euzenat, 2012). Common alignment methods used in state-of-the-art matching systems, based on similarity measurements, include:

- **Terminological**; comparison of the labels of entities.
- **Structural**; including internal comparison of entities and external comparison of relations with other entities.
- **Extensional**; analyzing the data instances in the ontology.
- **Semantic**; comparing the models of the entities.

While many systems such as those from Seddiqui and Aono (2009) and Cruz et al. (2009) use combinations of terminological and structural methods, the use of extensional and semantic methods in systems such as the one by Jean-Mary et al. (2009) have been largely unexplored (Pavel and Euzenat, 2012).

Similarly to these approaches, we use a combination of alignment methods to create mappings between PropBank (Palmer et al., 2005) predicates and DBpedia (Auer et al., 2007) properties. In particular, we identify predicate–argument structures from Wikipedia articles to extract triples and use a combination approach of extensional and semantical methods during the matching process to align the extracted triples with an exisiting DBpedia dataset.

2. System Description

Our system consists of different modules that perform the text processing tasks in parallel. Taking a set of Wikipedia articles as input, it produces PropBank-DBpedia ontology mappings. A generic semantic processing component based on a semantic role labeler (SRL) identifies the relations in the Wikipedia article texts. A coreference resolution module detects and links coreferring mentions in text and uses them to link the mentions located in the arguments of relations. Using a named entity linking module together with information inferred from the coreference chains, mentions are linked to a corresponding DBpedia URI. Finally, an ontology mapping module performs the final mapping of predicates from the PropBank nomenclature onto the DBpedia namespace.

3. Method

Using PropBank as a dictionary, our semantic parser annotates sentences with predicate–argument structures called rolesets. Our goal is to map more than 7,000 rolesets defined by PropBank, onto a more generalized roleset described by 1,650 DBpedia properties.

The matching process consists of the following steps:

1. Given a set of n-ary predicate–argument relations, we create binary subject–predicate–object relations by combinatorial generation.
2. The subject and object of the extracted relations are matched exactly to existing triples in the DBpedia dataset.
3. From the matching set of triples, links between PropBank roles and DBpedia properties are created. The mappings with the highest counts are selected.

We create then a generalized set of mappings using two procedures:

1. We generalize the subjects and objects of the extracted triples containing DBpedia URIs to 43 top-level DBpedia ontology classes.
2. We generalize the objects containing strings, dates, and numbers to the categories: String, Date, and Number respectively.

Most systems express mappings as alignments between single entities belonging to different ontologies. In addition, we also retain the related subject and object entities in such alignments and use them to express a more detailed mapping.

4. Results and Evaluation

In total, we processed 114,895 articles and we extracted 1,023,316 triples. The system mapped successfully 189,610 triples mapped to the DBpedia ontology. We singled out the unmapped triples using a predicate localized to a custom PropBank namespace. In Table 1, we can see that from the 189,610 extracted triples, 15,067 triples already exist in the DBpedia dataset. This means that our framework rediscovered 15,067 triples during the matching phase and introduced 174,543 new triples to the DBpedia namespace.

Subject	Predicate	Object	Mapping
dbpedia-owl:Person	bear.02.AM-LOC	dbpedia-owl:Place	dbpedia-owl:birthPlace
dbpedia-owl:Person	bear.02.AM-TMP	Date	dbpedia-owl:birthDate
dbpedia-owl:Person	retire.01.AM-TMP	Numeric	dbpedia-owl:activeYearsEndYear
dbpedia-owl:Person	marry.01.A1	dbpedia-owl:Person	dbpedia-owl:spouse
dbpedia-owl:Person	receive.01.A1	Thing	dbpedia-owl:award
dbpedia-owl:Person	manage.01.A1	dbpedia-owl:Organisation	dbpedia-owl:managerClub
dbpedia-owl:Person	serve.01.A1	dbpedia-owl:Organisation	dbpedia-owl:militaryBranch
dbpedia-owl:Place	locate.01.AM-LOC	dbpedia-owl:Place	dbpedia-owl:isPartOf
dbpedia-owl:Place	open.01.AM-TMP	Date	dbpedia-owl:openingDate
dbpedia-owl:Place	build.01.AM-TMP	Numeric	dbpedia-owl:yearOfConstruction
dbpedia-owl:Place	lie.01.A2	dbpedia-owl:Place	dbpedia-owl:locatedInArea
dbpedia-owl:Place	region.01.A1	dbpedia-owl:Place	dbpedia-owl:region
dbpedia-owl:Place	base.01.AM-LOC	dbpedia-owl:Place	dbpedia-owl:capital
dbpedia-owl:Place	include.01.A2	dbpedia-owl:Place	dbpedia-owl:largestCity
dbpedia-owl:Organisation	establish.01.AM-TMP	Numeric	dbpedia-owl:foundingYear
dbpedia-owl:Organisation	find.01.AM-TMP	Date	dbpedia-owl:formationDate
dbpedia-owl:Organisation	base.01.AM-LOC	dbpedia-owl:Place	dbpedia-owl:location
dbpedia-owl:Organisation	serve.01.A2	dbpedia-owl:Place	dbpedia-owl:broadcastArea
dbpedia-owl:Organisation	own.01.A1	dbpedia-owl:Organisation	dbpedia-owl:subsidiary
dbpedia-owl:Organisation	provide.01.A1	Thing	dbpedia-owl:product
dbpedia-owl:Organisation	include.01.A2	dbpedia-owl:Person	dbpedia-owl:bandMember

Table 2: Twenty one of the most frequent ontology mappings learned through bootstrapping.

Type	Count
DBpedia mapped triples (of which 15,067 already exist in DBpedia)	189,610
Unmapped triples	833,706
Total	1,023,316

Table 1: The extracted triples.

Table 2 shows some of the most frequent mappings learned during the matching process. Some general mappings, such as a person marrying another person corresponding to a spouse property, may hold for all cases. However, a mapping describing a person receiving a thing corresponding to an award property, requires a more detailed analysis since the thing received may represent items other than awards. Therefore, we believe that our ontology matching can be improved by a more fine grained approach to the subject-object generalization. In addition, by utilizing interlinking between DBpedia and other datasets such as LinkedMDB[1], we believe we can increase the amount of bootstrapping instances and thereby create more finely expressed mappings of higher quality.

5. Conclusion and Future work

From more than 114,000 articles from the English edition of Wikipedia, we have created a set of mappings aligning PropBank rolesets to DBpedia properties. We have expressed the mappings as a set of links from subject–predicate–object relations to DBpedia properties. In addition, the mappings have been generalized by classifying entities in subjects and objects to 43 top–level DBpedia classes.

We will improve this work by utilizing a more fine-grained approach to generalization by making full use of over 320 DBpedia classes as expressed by the ontology. We also intend to improve the matching process by interlinking related datasets, thereby increasing the amount of training instances used for alignment.

6. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin / Heidelberg.

Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. 2009. Agreementmaker: efficient matching for large real-world schemas and ontologies. *Proc. VLDB Endow.*, 2(2):1586–1589, August.

Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabuka. 2009. Ontology matching with semantic verification. *Web Semantics*, 7(3):235 – 251.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Shvaiko Pavel and Jerome Euzenat. 2012. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints):1.

Md. Hanif Seddiqui and Masaki Aono. 2009. An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(4):344 – 356.

[1] http://www.linkedmdb.org/

Using the probability of readability to order Swedish texts

Johan Falkenjack[1], **Katarina Heimann Mühlenbock**[2]

(1) Santa Anna IT Research Institute AB, Linköping, Sweden
(2) Språkbanken, University of Gothenburg, Gothenburg
johsj@ida.liu.se, katarina.heimann.muhlenbock@gu.se

Abstract

In this study we present a new approach to rank readability in Swedish texts based on lexical, morpho-syntactic and syntactic analysis of text as well as machine learning. The basic premise and theory is presented as well as a small experiment testing the feasibility, but not actual performance, of the approach. The experiment shows that it is possible to implement a system based on the approach, however, the actual performance of such a system has not been evaluated as the necessary resources for such an evaluation does not yet exist for Swedish. The experiment also shows that a classifier based on the aforementioned linguistic analysis, on our limited test set, outperforms classifiers based on established metrics used to assess readability such as LIX, OVIX and Nominal Ratio.

1. Motivation

Studies have shown that as many as 25 % of the Swedish adult population can not read at the level expected of students in the 9th grade in the Swedish school system. For many of these people, access to information is dependent on the ability to find the most easy-to-read texts describing the subject.

To this purpose a search engine, Webblättlast, capable of finding not only the most relevant texts but also the most easy-to-read is being developed at the Department of Computer and Information Science at Linköping University. This search engine mainly uses the three established Swedish metrics, LIX (Läsbarhetsindex (Björnsson, 1968)) which is a readability metric based on surface structure and OVIX (Ordvariationsindex (Hultman and Westman, 1977)) and Nominal ratio (Hultman and Westman, 1977) which are complexity metrics which measure word variation and information density respectively.

However, research has shown that these established Swedish readability metrics are insufficient when used individually (Mühlenbock and Johansson Kokkinakis, 2009). Also, the same study showed that LIX and OVIX very well might result in different orderings when used to rank documents according to supposed degree of readability.

2. Background

The years since 2000 have seen quite a few developments in the field of readability assessment for English. Some new readability assessment systems have utilized tools such as grammar parsers and discourse analysis (Feng et al., 2009). Other more data intensive studies have applied statistical language models such as n-gram models to the field of readability assessment with good results (Collins-Thompson and Callan, 2004). Most of these approaches were based on access to a corpus of readability assessed texts, Weekly Reader.

All texts in the Weekly Reader corpus are tagged with a suitable grade level in the U.S. school system. These grade levels can been used both as a basis for regression (Pitler and Nenkova, 2008) and for creation of detectors, single-class classifiers (Petersen, 2007). Both a formula generated by regression and a set of detectors can be used for ranking documents according to degree of readability.

However, no equivalent corpora exist for Swedish so another approach must be devised.

3. A new approach

If the assumption is made that *the degree of readability of a text is proportionate to the probability that the text is classified as easy-to-read by a perfect classifier* the problem becomes one of constructing such a classifier and finding a way to extract probabilities from it. Of course, such a perfect classifier is a purely theoretical construct. However, a good enough linear classifier, tweaked to output class probabilities (soft classification) rather than just the most probable class (hard classification), should be able to calculate a reasonable approximation, at least within a limited span. While this metric might not be linear, and therefore perhaps not suitable for single document assessment without some kind of smoothing, it should provide a way to rank documents based on their degree of readability.

4. Feasibility of the approach

To test whether the approach is feasible we first have to test whether a traditional classifier, able to identify easy-to-read texts, can be constructed. To do this we used documents from the LäSBarT easy-to-read corpus (Mühlenbock, 2008) to represent easy-to-read texts and documents from the GP2007, a corpus made up of articles from the newspaper Göteborgsposten from 2007, to represent non-easy-to-read texts.

4.1 Hard classification

These documents were analysed using the Korp corpus import tool developed by Språkbanken and six different models were created. Three models were based on the established Swedish readability metrics LIX, OVIX and Nominal ratio (NR), the fourth model (COM) combined all three, the fifth model (NODEP) added further surface, lexical and morpho-syntactic features and the sixth and last model (FULL) added features based on dependency parsing. These larger models are similar to the ones used by the Italian READ-IT project (Dell'Orletta et al., 2011).

Using the Waikato Environment for Knowledge Analysis, or WEKA, we tested the accuracy of a support vector machine (using the sequential minimal optimization training algorithm (Platt, 1998)) with the six different models. Each model was tested using 7-fold cross-validation over 1400 documents, 700 from each corpus. See Table 1 for the results.

Model	Accuracy
LIX	77.4
OVIX	84.5
NR	53.0
COM	89.3
NODEP	97.0
FULL	97.6

Table 1: The accuracy, percentage of correctly classified documents, for each model using hard classification.

Overall results, a maximum accuracy of 97.6 %, implies that it is possible to create a reasonably accurate classifier for easy-to-read Swedish documents.

4.2 Soft classification

If we are to order documents based on their probability of readability our classifier must be able to output a class probabilities. This can, for our classifier, be done by fitting logistic models to the output from the SVM. (Another approach would be to use logistic regression alone but as an SVM was the most accurate classifier in a related experiment we decided to use this hybrid approach.)

We must, however, make sure that this does not impair the accuracy of the classifier. We should also check the number of equivalence classes generated in the test, that is, how many documents are awarded the same probability of readability and therefore not sortable with regard to each other. As only the NODEP and the FULL models had accuracies $>$ 90 % only these models were evaluated. All documents with an error smaller than 50 % were considered correct. The same 7-fold cross validation scheme was used again and the calculated percentages were registered to calculate the number of equivalence classes. See Table 2 for the results.

Model	Accuracy	#Equivalence classes
NODEP	97.7	1398
FULL	97.0	1398

Table 2: The accuracy and number of equivalence classes for each model using soft classification.

The result shows that the accuracy is still high and the number of equivalence classes shows that all but 3 documents are sortable.

5. Conclusion and future work

The experiment shows that a SVM classifier with a high accuracy on readability based classification also can order documents without a large risk of non-sortable pairs. This implies that a system for ranking documents based on this principle could be feasible. However, other classification algorithms, more suited for what we call soft classification, such as pure logistic regression, might be more effective as they might produce more normalized results.

However, until tested we can not be sure that the ability to accurately identify easy-to-read documents entails the ability to order documents according to degree of readability. Further research, using ordered sets of documents, is necessary.

6. References

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Kevyn Collins-Thompson and Jamie Callan. 2004. A Language Modeling Approach to Predicting Reading Difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, July.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel, Lund.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited - An extended readability measure. In Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith, editors, *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, UK, July 20-23.

Katarina Mühlenbock. 2008. Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In Anju Saxena and Åke Viberg, editors, *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Sarah Petersen. 2007. *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, University of Washington, Seattle, WA.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, HI, October.

John C. Platt. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research, April.

Practical aspects of transferring the English Berkeley FrameNet to other languages

Karin Friberg Heppin, Kaarlo Voionmaa

Språkbanken, Department of Swedish
University of Gothenburg
karin.friberg.heppin@svenska.gu.se, kaarlo.voionmaa@svenska.gu.se

Abstract

This paper discusses work on annotating data in semantic frames. The work is carried out in the Swedish Framenet project, SweFN++, and it is theoretically and practically closely connected to the English Berkeley Framenet. In order for the framenet to be a useful tool for linguistic research and for applications based on it, it is essential that problems encountered while annotating the data will be thoroughly discussed before any solutions are provided. Here, two questions are in focus: firstly, the problematic relation between certain word combinations in English that translate as compounds in Swedish, and secondly, the cross-language semantic differences between English and Swedish. These problematic cases are brought to discussion in order to find solutions to them at a later phase. Languages differ, and it is crucial for the framenet endeavour to describe and analyse these differences in a systematic and valid manner drawing theoretically on the frame semantics.

1. Introduction

In what follows, we first briefly describe the theoretical and practical basics of framenet. We then proceed to describe some problems and difficulties. Concluding the paper, we emphasize the importance of treating every language on its own terms and not adjusting too much to the Berkeley FrameNet.

2. Framenet in various languages

The theoretical approach of framenet is based on frame semantics, brought forward by Charles Fillmore (Fillmore et al., 2003; Ruppenhofer et al., 2010). A lexical unit (LU) in framenet is a pairing of a word or multiword expression with its meaning. Each sense of the word or expression evokes a different frame, which is a script-like conceptual structure describing a type of situation, object or event along with typical participants described in terms of semantic roles. These participants are called frame elements.

The concepts behind the framenet frames are in principle language independent, while the lexical units and the annotated sentences are language dependent. It is this duality that makes cross-language comparisons of more complete framenet versions fruitful. Linking frames of different languages make concept equivalents visible, and studying corresponding sets of frames closer illustrate the differences. This could be useful, for example in various pedagogical and didactic applications in the area of second language teaching and learning. Framenet may provide a model for describing cross-linguistic similarities and differences through the relations of the conceptual frames, the language specific lexical units, and the syntax of the annotated sentences (Friberg Heppin and Friberg, 2012).

The English BFN[1], constructed by the Berkeley research team, is the basis for framenets in various languages such as the SweFN.[2] A fairly up-to-date treatment of framenet analyses on various languages is found in Boas (2009).

BFN contains presently more than 10,000 lexical units in more than 1,000 frames, together with more than 170,000 sentences. The Swedish framenet project started in 2009 and turned into a full scale project in 2011 (Borin et al., 2010; Friberg Heppin and Toporowska Gronostaj, 2012). By October 2012, SweFN covered 684 frames comprising around 23,200 lexical units. Each lexical unit is a unique entry in SALDO, a Swedish large-scale semantic lexical resource for language technology (Borin et al., 2008).

3. Problems and difficulties

It should be noted that in annotating frames in SweFN, we follow the BFN procedures as regards frame names and the selection of frame elements including their definitions and internal relations. Even meta-information about the frames, such as the semantic relations between the frames, is transferred to SweFN.

A specific area of concern for the Swedish framenet team has proved to be the semantic features of compound lexical units as the process of forming orthographic compounds is very productive in Swedish. To exemplify: there are a number of LUs in the Berkeley FrameNet the equivalents of which do not occur as separate words in Swedish, but only as compound constituents. This is the case with *key* and *main* from the Importance frame. The Swedish equivalences *nyckel-* and *huvud-* occur as prefixes in compounds such as *nyckel*|*fråga* ‘key question’ and *huvud*|*gata* ‘main street’. Appearing as individual words they have the concrete sense of ‘key’ and ‘head’ and not the prefix senses: ‘most important’ and ‘most central’ respectively.

Related is the case where the head of the compound is very rare as a simplex word. One may consider the Swedish word *brott* in the sense ‘quarry’, which evokes the frame Mining. This LU is so uncommon that it presently has no entry in SALDO. However, there are compounds in SALDO with *brott* as head, such as *dag*|*brott* ‘opencast mine’ and *sten*|*brott* ‘quarry’. This puts into question the

[1] <https://framenet.icsi.berkeley.edu/fndrupal/>

[2] <http://spraakbanken.gu.se/eng/swefn> SweFN is available as a free resource (CC-BY-SA 3.0, LGPL 3.0)

guideline of the Swedish framenet according to which the head of a compound, as a simplex word, should evoke the same frame as the compound itself.

The polysemous word *brott* can further illustrate the problem of treating compounds. Looking now at *brott* in the sense of 'fracture' which could be a break in a long entity, such as a cable or pipe, but in a more specific sense it could be a bone fracture. There is, at the moment, no existing frame in the BFN which would be evoked by the general meaning of this noun, only related frames such as: Breaking_apart and Becoming_separated. However, there are frames in both SweFN and BFN that could be evoked by the LU in the sense 'bone fracture'. In BFN it is the frame Medical_conditions which in SweFN has been replaced with the more specific frames Health_status and Medical_disorders. The question is if it is motivated to conclude that *brott* 'fracture' has two separate senses and thus should be denoted as two separate LUs, one in a new frame for the general sense and one in Medical_conditions/Medical_disorders for 'bone fracture'. If the decision is to conclude that *brott* 'fracture' is one LU only, it should populate the as of yet non-existing frame which might be called Fracture. However, this would entail that compounds such as *benbrott* 'bone fracture', which evoke the frames Medical_conditions/Medical_disorders, would not populate the same frame as the head of the compound: *brott*.

Another problem concerns the lists of lexical units populating the frames. These lists are language dependent. The difference is not only in the LUs themselves, but also in the fact that some frames may only have periferal and infrequent LUs in a certain language, that do not cover all aspects of the frame. The concept may be expressed through constructions, such as multiword patterns, rather than by lexical units. This is the case with the English frame Indicating which contains a single lexical unit: *name* (verb). There are LUs in Swedish which evoke the frame, such as *räkna upp*, lit. 'count up', which is only used to refer to lists of entities or *namnge*, lit. 'give name', which must refer to a proper name. In other cases, such as in an equivalent of the English phrase 'Name that sound' special constructions must be used.

The frame Indicating illustrates a risk in building a framenet with startingpoint in another language. A frame can in one language contain LUs which evidently belong there, while for another language there are only farfetched unusual ones, thay may only pertain to certain jargongs. The concept might in such a case be better expressed by a certain construction and thus the frame and the contained LUs may be misleading, especially for second language learners. A solution may be to connect the framenet to a constructicon, a collection of schematic multi-word units or constructions. Constructicons based on framenet are presently being constructed, although still in early stages, for languages such as English (Fillmore et al., to appear) and Swedish (Lyngfelt et al., 2012).

4. Concluding remarks

Framenet is an endeavour which operates on a global level and on a very broad lexico-semantic base. Naturally, problems inevitably arise in a project of this size. Above, we have discussed some of these problems.

The point of departure for framenets in various languages has been the Berkeley FrameNet, and this will continue to play a major role. While the work proceeds with and in different languages, we will become ever more aware of points in common and differences. Squeezing problematic cases into the patterns provided by a master framenet we might lose crucial linguistic information, and thus, diminish the research value of the whole framenet endeavour.

5. Acknowledgements

The research presented here was supported by the Swedish Research Council (the project Swedish Framenet++, VR dnr 2010-6013) and by the University of Gothenburg through its support of the Centre for Language Technology and Språkbanken (the Swedish Language Bank).

6. References

Hans C. Boas, editor. 2009. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Mouton de Gruyter.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a freely available lexical database for swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beáta Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, Studia Linguistica Upsaliensia, pages 21–32. Acta Universitatis Upsaliensis.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in Swedish Framenet++. <https://svn.spraakdata.gu.se/sb/fnplusplus/pub/SweFN_Euralex_extended.pdf>.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L.Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3).

Charles Fillmore, Russel Lee Goldman, and Russell Rhodes, to appear. *Sign-Based Construction Grammar*, chapter The FrameNet Constructicon. CSLI, Stanford.

Karin Friberg Heppin and Håkan Friberg. 2012. Using FrameNet in communicative language teaching. In *Proceedings of the XV EURALEX International Congress*, Oslo, Norway.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2012. The rocky road towards a Swedish FrameNet. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.

Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Adding a constructicon to the Swedish resorce network of Språkbanken. In *Proceedings of KONVENS 2012*. LexSem workshop.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended theory and practice*. <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>.

Translating English texts into sets of predicate argument structures

Elisabeth Godbert, Jean Royauté

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
CNRS UMR 7279 - Aix-Marseille Université
Parc Scientifique et Technologique de Luminy, case 901
13288 Marseille Cedex 9
Elisabeth.Godbert@lif.univ-mrs.fr, Jean.Royaute@lif.univ-mrs.fr

1. Introduction

This paper focuses on predicate argument structures (PAS) in English texts and on the representation of events and states described in them. An event is an action either in progress or achieved. A state is a property of an object. In most cases, verbs denote actions and adjectives denote states.

Verbs and adjectives (used for example with *be*) are roots of sentences. Their nominalizations, whose syntactic patterns are more difficult to parse, allow to express the same information, except for tense marks.

We study five types of predicates: verbs, adjectives, nominalizations of verbs, nominalizations of adjectives and predicate nouns which are not related to a verb.

In most cases, verbs, adjectives and their nominalizations have the same argument relations, where arguments play precise semantic roles: they are core arguments (subjects/objects) or adjuncts. In noun phrases (NP) with nominalizations, the head noun is bound to prepositional phrases (PP) with specific prepositions which introduce arguments. For example, the NP *activation of T cells by dendritic cells* is related to the verbal form *dendritic cells activate T cells* and it is possible to insert an adjunct into the two frames, such as *in the context of pathogens*. Core arguments can also take the place of a modifier as in *T cells activation*.

Nominalizations are numerous in texts, but parsing NPs can be rather complex when they contain several nominalizations and PPs, and errors in parses are often due to incorrect prepositional attachments (Miyao et al., 2006). The first part of our work consists in a detailed study of all possible syntactic patterns for predicates, which will help us to improve prepositional attachments in parses.

The second part of our work is the translation of sentences (events and states they express) into sets of PAS expressed in an underspecified semantics. This semantics is based on three macro-roles: Agent (or Cause), Patient (or Theme) and Circumstance. The Agent is the argument which performs the action in the case of an event or to which is attached a property in the case of a state. The Patient is the argument which is involved by an action or by a state. In an active verbal form, the subject is the Agent and the object complement is the Patient, which can be introduced or not by a preposition. In passive form, roles are inverted. Circumstance is the third semantic role and corresponds to adjuncts. Thus, this underspecified semantics is at interface between syntax and semantics.

2. Typology of predicates

A typology of predicates has been defined, according to all their possible syntactic patterns. Then, predicates have been classified into seven main classes described in (Godbert and Royauté, 2010). This classification has been elaborated from scientific texts of the web, from a grammar of English and from the data of ”The Specialist Lexicon” (Browne et al., 2000). Two criteria have been used to define the seven classes: (i) the role of the preposition *of* in the NP, which can mark a subject or an object complement and (ii) the role of arguments of symetric predicates, for which arguments can be exchanged. Here are a few examples from the seven classes:

- Classes 1 and 2 group together verbs accepting a direct object and the passive voice.

Heat activates electrons / Activation of electrons by heat.

John attributes human emotions to animals / Attribution of human emotions to animals by John.

- Class 6 concerns predicates with interchangeable arguments: subject and object can permute without changing the meaning.

Genes Interact with proteins / Interaction of genes with proteins / Interaction of/between genes and proteins.

Lisbon Treaty is concordant with the Czech constitution / The concordance of the Lisbon Treaty with the Czech constitution.

3. The PredXtract system

The PredXtract system is based on the Link Parser (LP) and its English native Link Grammar (LG), a variant of dependency grammars (Sleator and Temperley, 1991).

Our domain of application is biomedical text, so we have added to LG a lexicon and grammar of biological terms. The lexicon contains about 500,000 inflected forms.

In LG, links that attach verbs or nouns to any prepositional complement are generic links. In order to improve prepositional attachment and to mark the precise role of each argument of predicates, we have defined specific argument links and integrated them into the grammar.

A new grammatical module, based on argument links, has been developed for nominalizations. At the conclusion of our study of all possible syntactic patterns of nominalizations, 110 subclasses have been defined within the seven main classes mentionned in Section 2. This nominalization module contains the syntactic features of about 7,350 nominalizations, splitted into the 110 subclasses. Each nominalization can accept one or more syntactic descriptions and thus can belong to several subclasses.

Besides, several modules have been developed for post-processing the parses produced (for one sentence, often several thousands parses are produced).
The verb-adjective-noun alignment module aligns verb and adjective arguments to nominalization arguments in all parses: it integrates argument links when appropriate, identifies each verbal (or adjectival) sequence (verb with possible auxiliaries and modalities), and it identifies arguments in passive or active voice, and interchangeable arguments.
Then, for each sentence, the parses are reordered by attributing to each parse a score defined through several criteria. These criteria mainly take into account argument links in parses. For example, in the case of multiple prepositional attachments, we favor (i.e. give a higher score to) parses whose number of argument links is maximum.
At last, the syntax-semantics interface module produces for each sentence its underspecified semantic representation, close to the syntax, expressed in terms of the three macro-roles Agent, Patient and Circumstance.

4. Parsing biomedical texts

An evaluation of PredXtract, for the identification of arguments of verbs and nominalizations of verbs, has been performed on a corpus of 400 random sentences from 3500 sentences of Medline abstracts. In the corpus, nominalizations represented 42.3% of all predicates. The system obtained rather good results for the identification of arguments: F-measure of approximatively 0.88 for true arguments but possibly not completely reconstituted, and 0.78 if only true and complete arguments were scored true.

Below are two examples of output of the system. Active/passive forms are noted A/P. In Ex.1 we can note (i) three nominalizations of verbs (*isolation, translation, growth*), and a nominalization of adjective (*importance*), (ii) the use of the modal *may* which operates on the verb *reflect* and is included in the verbal sequence and (iii) an error on the attachment of Circumstance with *during*, attached to *importance* instead of *involved*. According to our definition in Section 1., these PASs show one state (*importance*) and five events (*isolation, translation, growth, reflect, involved*). In Ex.2 we can note the two permutable arguments Agent A and B of *interaction*, the "That clause" of *propose* and the modal *may* on the verb *influence*.

```
======================================================
Ex.1: Isolation of P. temperata def may reflect the
importance of specific amino acids involved in the
translation process during growth in the insect host
------------------------------------------------------
Nominalization 1: isolation
   Patient: P. temperata def
Nominalization 2: importance
   Agent: specific amino acids involved in [..] process
   Circumstance: {during} growth in [..] host
Nominalization 3: translation
Nominalization 4: growth
   Circumstance: {in} the insect host
Verb 1: reflect (verb.seq: may reflect)(A)
   Agent: isolation of P.temperata def
   Patient: the importance of specific [..] process
Verb 2: involved (verb.seq: involved)(P)
   Patient: specific amino acids
   Patient: {in} the translation process
======================================================
```

```
Ex.2: We propose that aberrant interaction of mutant hunting-
tin with other proteins may influence disease progression
------------------------------------------------------
Nominalization 1: interaction
  Agent A: mutant huntingtin
  Agent B: other proteins
Nominalization 2: progression
  Patient: disease
Verb 1: propose (verb.seq: propose) (A)
  Agent: we
  That clause: that aberrant interaction [..] progression
Verb 2: influence (verb.seq: may influence) (A)
  Agent: aberrant interaction of [..] proteins
  Patient: disease progression
======================================================
```

Adaptation to BioNLP 2011 Shared Tasks
BioNLP 2011 Tasks aimed at fine-grained information extraction (IE) in the domain of biomolecular event extraction (Kim et al., 2011). For example, from the sentence *PmrB is required for activation of PmrA in response to mild acid pH*, two events (E1 and E2) must be extracted:
E1:Pos-regulation;activation;Theme:PmrA
E2:Pos-reg[..];required;Cause:PmrB;Theme:E1
Events are defined by trigger words (verbs, nouns, adjectives or complex expressions) and their arguments which are biological entities or other events. The main argument roles are "Cause" (Agent) and "Theme" (Patient).
We have participated to BioNLP Tasks. For that, we had to adapt PredXtract: complete it with specific modules for preprocessing data before parsing and for postprocessing output to extract events and write them in the right format.

5. Discussion

Much research has been done on PAS but it is difficult to compare them because objectives are often different (see for example (Johansson and Nugues, 2008) on WSJ). PredXtract includes the results of an extensive study of syntactic patterns of verbs, adjectives and nominalizations. Nominalizations are numerous in biomedical text but other research on nominalizations in biomedecine is very limited. PredXtract has been adapted in a short time to specific IE tasks for BioNLP. We now aim to use it in other domains.

6. References

A. C. Browne, A. T. Mccray, and S. Srinivasan. 2000. The specialist lexicon technical report. *Lister Hill National Center for Biomedical Communications, NLM, USA.*

E. Godbert and J. Royauté. 2010. Predxtract, a generic platform to extract in texts predicate argument structures. *Workshop "Semantic Relations", LREC 2010.*

R. Johansson and P. Nugues. 2008. Dependency-based semantic role labeling of propbank. *Proc. of the 2008 Conference on Empirical Methods in NLP*, pages 69–78.

J. Kim, S. Pyysalo, T. Ohta, R. Bossy, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. *BioNLP 2011 Workshop Volume for Shared Task, ACL*, pages 1–6.

Y. Miyao, O. Tomoko, M. Katsuya, T. Yoshimasa, Y. Kazuhiro, N. Takashi, and J. Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. *Proc. of the COLING-ACL 2006*, pages 1017–1024.

D.D. Sleator and D. Temperley. 1991. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196.*

Towards an automatic detection of the chemical risk statements

Natalia Grabar[1], **Laura Maxim**[2], **Thierry Hamon**[3]

(1) CNRS UMR 8163 STL, Université Lille 1&3, France
(2) Institut des Sciences de la Communication, CNRS UPS 3088, France
(3) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, France
natalia.grabar@univ-lille3.fr, laura.maxim@iscc.cnrs.fr, thierry.hamon@univ-paris13.fr

Abstract

We present an experiment on the detection of the chemical risk statements in institutional documents. The method relies on linguistic annotation and exploitation of classes, which describe the risk factors, and linguistic resources (negation, limitations and uncertainty markers). The method provides promising results. It will be enriched with more sophisticated NLP processing.

1. Introduction

Early detection of chemical risks (harmful effects of chemical substances on human health or the environment), such as those related to Bisphenol A or phtalates in the scientific and institutional literature may play an important role on the decisions made on marketing of the chemical products and has important concerns to the public health and security. Given the tremendous amount of the literature to be analyzed, it becomes important to provide automatic methods for the systematic mining of this literature.

2. Material and Methods

We work with four types of material: (1) classes which describe factors related to chemical risk, (2) document to process, (3) linguistic resources, and (4) reference data. Risk factor classes describe factors like causal relationship between the chemicals and the induced risk, laboratory procedures, human factors, animals tested, exposure, etc. Each class receives a short label, such as *Form of the dose-effect relationship*, *Performance of the measurement instruments* or *Sample contamination*. The processed document has been created by EFSA (European Food Safety Authority) in 2010. It proposes a literature review on Bisphenol A-related experiments and known risks or suspicions. It contains 116 pages and over 80,000 word occurrences. This is a typical institutional report which supports the decisions for managing the chemical risk. Linguistic resources contain markers for negation (Chapman et al., 2001) (*i.e.*, *no, not, neither, lack, absent, missing*, which indicate that a result has not been observed in a study, a study did not respect the norms, etc.), uncertainty (Périnet et al., 2011) (*i.e.*, *possible, hypothetical, should, can, may, usually*, which indicate doubts about the results of a study, their interpretation, significance, etc), and limitations (*i.e.*, *only, shortcoming, small, insufficient*, which indicate limits, such as small size of a sample, small number of tests or doses, etc.). The reference data is obtained thanks to a manual annotation by a specialist of chemical risk assessment: 284 segments are extracted to illustrate 34 risk factor classes.

Figure 1 presents the main steps of the method. Preprocessing is done with the Ogmios plateform[1] and provides linguistically normalized text and class labels (tokenized, POS-tagged and lemmatized (Schmid, 1994)). Then, the text is automatically annotated with the linguistic resources. During the postprocessing, we try to make a link between class labels and the text. For this, we combine information from the annotation with linguistic resources (computed in number of the corresponding markers) and the lexical intersection between the class labels and text segments (computed in percents). For instance, in the segment: *However, no specific measures were adopted to avoid sample contamination with free BPA during analytical procedures, which therefore cannot be excluded*, we find three limitation and negation markers (*however, no, cannot*), and all the words from the class label *Sample contamination*. We test several thresholds for these two values. The final step of the method is the evaluation against the reference data.

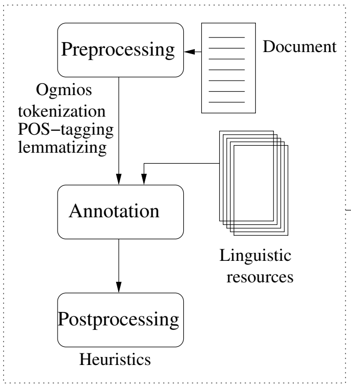


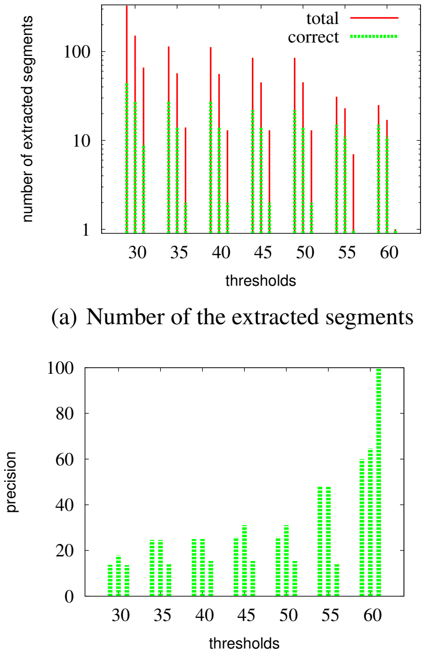
Figure 1: Main steps of the method.

3. Results and Discussion

On figure 2, we present the main results obtained. On the axis x, we indicate the applied thresholds (30%, 35%, 40% etc. of words in common), while the three impulses correspond to the presence of 1, 2 or 3 markers (limitation, negation, uncertainty). With the increasing of the constraints (number of markers and percentage of common words) the number of retrieved segments dicreases while the precision increases. The best thresholds seem to be 55% or 60%: the number of segments is then important, while their precision becomes acceptable (50-65%). Here are some examples:

1. Form of the dose - effect relationship: *There was no*

[1] *http://search.cpan.org/∼thhamon/Alvis-NLPPlatform-0.6/bin/ogmios-nlp-server*



(a) Number of the extracted segments

(b) Precision of the extracted segments

Figure 2: According to the tested thresholds: number of the extracted segments and their precision.

effect on testis weight in the BPA groups, and the lack of any dose response relationship in other organ weights does not suggest a treatment-related effect

2. Choice of the experimental unit, number of animal test simultaneously: *In addition, the study has some shortcomings (small experimental groups of 3-4 animals and evaluations in males only, no indication of the number of exposed dams, or whether animals in the tested groups were littermates)*

Among the 34 classes tested, the method currently detects segments for 18 classes. We performed also a manual analysis, which showed that the method detects also segments which are correct although they are not part of the reference data. If these segments were to be considered, the precision would increase by 10 to 15%. The manual analysis revealed also the current limitations of the method. For instance, in several extracted segments, there is no syntcatic nor semantic relation between the various markers and words from labels. To mend such extractions, a syntactic analysis should be exploited. Another limitation is when there is no direct correspondence between words used in the class labels and words used in the processed document, like in *GLP compliance* and *GLP compliant*. For this a specific lexicon of synonyms and morpho-syntactic variants will be developed. Otherwise, some labels may not be evocative of their full meaning or of the expressions used in the document: other methods will be designed for them. It remains difficult to compare this experience to the existing NLP work. The closest work is done in the project Met@risk (*http://www.paris.inra.fr/metarisk*), but up to now there is no published results. Otherwise, the risk management in other domains is tackled through the building of dedicated resources (Makki et al., 2008), exploring reports on known industrial incidents and searching for similar newly created documents (Tulechki and Tanguy, 2012), calculating the exposure (Marre et al., 2010) or information extraction (Hamon and Grabar, 2010).

4. Conclusions et Perspectives

We presented results of the first experiments performed in the automatic detection of the chemical risk statements. A set of specific classes describing the factors of the chemical risk is exploited. The labels of these classes together with negation, limitation and uncertainty markers are recognized in the processed institutional document and allow to extract segments which state about the chemical danger and insufficiency of current studies. With our best thresholds, the extracted segments show precision 50-65% which may be improved if the current reference data are completed. Up to now, the method is domain independent and relies only on the labels of the classes. But this method has to evolve: new functionalities (specific contextual rules) and resources (specific synonyms and morpho-syntactic variants) will be added in order to manage more risk classes and to explore the documents more exhaustively. In order to improve the precision, we will go beyond the cooccurrences and integrate the syntactic analysis and dependencies among the words. Moreover, the method will be applied to other regulatory and scientific risk assessment reports and studies, and to other substances. The extraction results will be anaysed with several experts of the chemical risk assessment.

Acknowledgement

This work is done within the frames of the AIR REACH and DicoRisque (11-MRES-PNRPE-4-CVS-30) projects.

5. References

WW Chapman, W Bridewell, P Hanbury, GF Cooper, and BG Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform. 2001 Oct;34(5):*, 34(5):301–10.

T Hamon and N Grabar. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, 17(5):549–54.

J Makki, AM Alquier, and V Prince. 2008. Ontology population via NLP techniques in risk management. In *Proceedings of ICSWE*.

A Marre, S Biver, M Baies, C Defreneix, and C Aventin. 2010. Gestion des risques en radiothérapie. *Radiothérapie*, 724:55–61.

A Périnet, N Grabar, and T Hamon. 2011. Identification des assertions dans les textes médicaux: application à la relation {patient, problàme médical}. *TAL*, 52(1):97–132.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, pages 44–49, Manchester, UK.

N Tulechki and L Tanguy. 2012. Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. In *TALN*, pages 439–446.

Linking expert and lay knowledge with distributional and synonymy resources: application to the mining of Diabetes Fora

Thierry Hamon[1], **Rémi Gagnayre**[2]

[1]Laboratoire d'Informatique Médicale et de BioInformatique (EA3969)
[2]Laboratoire de Pédagogie de la Santé (EA 3412)
Université Paris 13, Sorbonne Paris Cité, Bobigny, France
[1]thierry.hamon@univ-paris13.fr, [2]remi.gagnayre@univ-paris13.fr

Abstract

Online discussion fora are a valuable source for getting a better knowledge of patient skills and behaviours and then for improving therapeutic education programs. In this context, mining such forum messages requires to link patient vocabulary with expert description of the skills. To achieve it, we propose a method based on the exploitation of synonymy and distributional resources to enrich skill description before performing a mapping with forum messages. The proposed approach has a good coverage when the skills are lexically expressed. The results also show that such resources are helpful and complementary to identify skills.

1. Introduction

The knowledge on patient skills, i.e. the patient ability to understand and to daily manage his pathology, is crucial for healthcare professionals to create therapeutic education programs. Patient surveys are usually performed with potential bias of healthcare professional interference. Besides, the online discussion fora (ODF) are becoming a valuable source of information about the patient exchanges and behaviour (Dickerson et al., 2004). It is necessary to mine them but first the lay vocabulary of the patients must be linked with the expert description of the skills. Related works usually address the health literacy (McCray, 2005): the specialised term explanation (Liang et al., 2011), the definition of patient vocabularies (Zeng and Tse, 2006; Deléger and Zweigenbaum, 2008) or the identification of the specialisation level of texts (Poprat et al., 2006; Chmielik and Grabar, 2009). Few works focus on the study of the patient skills through patient exchanges: Chan et al. (2009) characterised exchanges between patients and physicians, Fredriksen et al. (2008) manually analysed a small amount of ODF messages concerning pregnancy problems. Previously, we could associate manually two thirds of ODF messages with a taxonomy describing the therapeutic education objectives (Harry et al., 2008). In this work, we aim at associating patient skills (d'Ivernois et al., 2011) described by experts with ODF messages written by patients, while both of them have their own vocabulary.

2. Material

Corpora We collect two corpora of ODF messages in French: *Les diabétiques*[1] (839 threads and 6,982 messages – 624,571 words); *Diabetes Doctissimo*[2] (22,532 threads and 560,066 messages – 35,059,868 words).

Skill descriptions A skill taxonomy has been defined for diabetic patient education (d'Ivernois et al., 2011). We used the 174 skill descriptions distributed in nine categories and composed of verbal phrases: *express the needs regarding the pathology*.

3. Lexical mapping of skills with messages

As the vocabulary used in the skill descriptions differs from the patient vocabulary, we have to increase the coverage and add a certain vagueness in the expanded descriptions. Thus, we enrich them linguistically and semantically by using 149,309 synonyms (between 48,859 words) of the French dictionary *Le Robert* and distributional resources providing associative relations: *FreDist*[3] (1,853,475 neighbours of 24,749 words), *Les voisins de Le Monde* (**VdLM**)[4] (2,762,739 neighbours of 145,164 words) and *Les voisins de Wikipedia* (**VdW**)[5] (43,690 neighbours of 173,853 words). These three distributional resources differ as for their source texts (French Wikipedia, newspaper *L'Est Républicain* or *Le Monde*) and methodology used for computing the distributions.

Skills are pre-processed: words are part-of-speech tagged and lemmatised, terms are recognised. Content words (nouns, adjectives and verbs) and terms extracted from the skill descriptions are considered as keywords and are expanded with each linguistic resource. The keywords and expanded forms are then mapped with words and terms from message corpora. Then the messages containing all the keywords associated to a given skill description or their expanded forms are selected and associated to this skill.

4. Experiments and Results

As baseline, we built keyword sets by only considering the content words. The mean number of keywords per skill is 3.4 for a total of 591 keywords. A keyword is shared by a 2.14 skill description on average.

Keyword expansion We defined several expanded keyword sets which include terms and one of the linguistic resources presented above. When adding terms, the mean number of keywords per skill increases to 4.02 and the set contains 701 keywords. The exploitation of the resources increases even more the size of the expanded keyword sets: on average, 23.58 synonyms and up to 673.11

[1] http://www.lesdiabetiques.com/modules.php?name=Forums

[2] http://forum.doctissimo.fr/sante/diabete/liste_sujet-1.htm

[3] http://fredist.gforge.inria.fr/

[4] http://redac.univ-tlse2.fr/applications/vdlm.html

[5] http://redac.univ-tlse2.fr/applications/vdw.html

VdLM neighbour words are associated to an expanded keyword. The linguistic resources help to expand the expert vocabulary but their contribution is variable: the vocabulary increases by ten times with the synonymy resource, and by twenty times for the **VdLM** and **FreDist** resources. Also, **VdW** leads to a vocabulary half as big as the synonymy resource. But the use of linguistic resource also increases the ambiguity in the skill description: a synonymous word is used to expand a mean of 3.8 keywords, and a **VdLM** neighbour word expands a mean of 66.23 keywords.

Skill mapping in corpora Messages are linguistically processed in the same way as the skill description. The skill mapping leads to select about 35% of the threads with the baseline. Unsurprisingly, terms reduce the number of retrieved threads and messages, while the linguistic resources increase this number (94 to 99% of the threads). But, at the message level, these observations vary according to corpora: on *LesDiabétiques*, only 7.5% of the messages are selected with the baseline, while we reach 70% of the messages when using synonyms, 85% with **freDist** and **VdW**, and up 93.4% with **VdLM**. The results are quite different on the corpus *Doctissimo*: only 3% of the messages are extracted with the baseline, 41% with synonyms and 50 to 66% with the distributional resources. The skills are not identified homogeneously: the skills belonging to the category *express the needs regarding the pathology* are retrieved in both corpora while skills from the categories *management of an emergency situation* and *metacognition, anticipate and plan the actions, and self-evaluation* are associated with no messages. Skill coverage variation can be due to unmentioned skills in the messages or, when mentioned, not expressed with the expected words. The resources have a variable influence on the skill coverage: synonymy identifies more skills than any other resources, while more messages are selected with the distributional resources. The use of terms has only an influence on the results from *Doctissimo*.

Manual analysis A random number of the selected messages are manually evaluated by an expert in health pedagogy and therapeutic education. The self-care skills are the most present in the messages: numerous discussions concern the treatment procedures and especially the insulin pump setup and its daily use. We also note that those are always mentioned with psychosocial skills. Asserting the health choices and rights are also among the frequent skills.

The approach faces difficulties when the skills are too abstract. In this situation, other external resources may be useful to improve the coverage and precision of the message selection. Also, in some cases, distributional resources are too ambiguous or semantically too far from the initial keyword meaning. We plan to limit the number of neighbours for a given word or to take into account the association strength between them.

5. Conclusion

The proposed approach has a good coverage when the skill keywords can be found in ODF: more messages are selected with the distributional resources while synonymy resource leads to the identification of more skills. The synonymy and distributional resources appear to be useful to link expert and patient vocabularies and to mine ODF. The results also confirm that it is possible to automatically identify skills daily used or mentioned by the patients from the online discussions. A manual analysis of a part of the selected messages shows that an important number of these messages is about the daily self-care and psychosocial skills. Moreover, the use of general synonymy dictionary and automatically built distributional resource ensure that our method is domain independent.

Several improvements can be done during the keywords expansion: exploitation of hierarchical relations to specialize some parts of the skills, specific processing of the thread subject, or removing the signature. Also, as the health ODF may contain some specific vocabulary or semantics non covered by the used resources, we plan to build additional resources from the ODF corpora.

Acknowledgments This work is funded by the Université Paris 13 (project BQR Bonus Quality Research, 2011).

6. References

C. V. Chan, L. A. Matthews, and D. R. Kaufman. 2009. A taxonomy characterizing complexity of consumer ehealth literacy. *Proceedings of the AMIA Symposium*, pages 86–90.

J. Chmielik and N. Grabar. 2009. Comparative study between expert and non-expert biomedical writings: their morphology and semantics. *Stud Health Technol Inform.*, 150:359–63.

L. Deléger and P. Zweigenbaum. 2008. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *Proceedings of the AMIA symposium*, pages 146–150.

S. Dickerson, A. M. Reinhart, T. Hugh Feeley, R. Bidani, E. Rich, V. K. Garg, and C. O. Hershey. 2004. Patient internet use for health information at three urban primary care clinics. *JAMIA*, 11(6):499–504.

J.-F. d'Ivernois, R. Gagnayre, and *et al.* 2011. Compétences d'adaptation à la maladie du patient : une proposition. *Educ Ther Patient/Ther Patient Educ*, 3(2):S201–S205.

E. Haukeland Fredriksen, K. M. Moland, and J. Sundby. 2008. "Listen to your body". A qualitative text analysis of internet discussions related to pregnancy health and pelvic girdle pain in pregnancy. *Patient Education and Counseling*, 73(2):294–9.

I. Harry, R. Gagnayre, and J.-F. d'Ivernois. 2008. Analyse des échanges écrits entre patients diabétiques sur les forums de discussion. *Distances et savoirs*, 6(3).

S. F. Liang, D. Scott, R. Stevens, and A. Rector. 2011. Unlocking medical ontologies for non-ontology experts. In *Proceedings of BioNLP Workshop*, pages 174–181, June.

A. T. McCray. 2005. Promoting health literacy. *JAMIA*, 12(2):152–63.

M. Poprat, K. Markó, and U. Hahn. 2006. A language classifier that automatically divides medical documents for experts and health care consumers. In *Proceedings of MIE*, pages 503–508.

Q. T. Zeng and T. Tse. 2006. Exploring and developing consumer health vocabularies. *JAMIA*, 13(1):24–29.

Nature Identical Prosody – data-driven prosodic feature assignment for diphone synthesis

Peter Juel Henrichsen
Danish Center for Applied Speech Technology (DanCAST)
Copenhagen Business School, Denmark
pjh.ibc@cbs.dk

Abstract

Today's synthetic voices are largely based on *diphone synthesis* (DiSyn) and *unit selection synthesis* (UnitSyn). In most DiSyn systems, prosodic envelopes are generated with formal models while UnitSyn systems refer to extensive, highly indexed sound databases. Each approach has its drawbacks; such as low naturalness (DiSyn) and dependence on huge amounts of background data (UnitSyn). We present a hybrid model based on high-level speech data. As preliminary tests show, prosodic models combining DiSyn style at the phone level with UnitSyn style at the supra-segmental levels may approach UnitSyn quality on a DiSyn footprint. Our test data are Danish, but our algorithm is language neutral.

1. Introduction

We outline a new method for improving the prosodic quality of artificial voices based on concatenative synthesis, inheriting the perceived naturalness of the massively data-demanding unit selection synthesis (UnitSyn) while maintaining the rational design of the conceptually simpler diphone synthesis (DiSyn).

The DiSyn engine is based on a sound database of a highly systematic design. The database can be described as a matrix *P*x*P*, where *P* is the phone inventory of the target language *T*. Each cell in the matrix is inhabited by a sound file representing a diphone (excluding those never occurring in *T*). Synthesis, then, amounts to diphone-splicing and post-processing. Due to the rational layout of the database, the footprint of the DiSyn system is moderate.

In the UnitSyn engine, in contrast, parsimony is traded for naturalness by including (huge amounts of) samples of connected speech in the database. Input text to the UnitSyn system with exact matches in the database are reproduced flawlessly (resembling playback rather than resynthesis), naturalness declining gracefully with the distance between input and best database match. In practical use UnitSyn systems tend to fluctuate between playback quality (very high) and sub-DiSyn quality (poor). In contrast, DiSyn systems deliver a moderate, but far more consistent quality.

	DiSyn (diphone)	**UnitSys (unit selection)**	**NIP (hybrid)**
Database preparation	*limited*	*labour-intensive*	*moderate*
Footprint	*moderate*	*very large*	*moderate*
Naturalness	*low*	*medium*	*medium*
Consistency	*high*	*low*	*high*

Table 1. Prosodic models (concatenative synthesis)

Our NIP algorithm (Nature Identical Prosody) combines the compact design of the DiSyn database with the data-driven prosodic plasticity of the UnitSyn. NIP can be applied in existing DiSyn systems, in contrast to other recently suggested hybrid synthesis systems (e.g. Oparin 2008, Aylett 2008, Guner 2011).

We first introduce Grønnum's prosodic model for the Danish sentence as well as our data-driven alternative; then we report on an experiment showing that a DiSyn-style algorithm informed by speech data may approach the UnitSyn prosodic quality.

2. Theory-driven prosody assignment

Following Grønnum (1978, 1985, 1992, 1998), the Danish stress group (SG) consists of one or more syllables. The rules of prosody assignment are:

I. an initial stressed syllable (all others unstressed),
II. from I, an F0 upstep to the 2nd syllable,
III. from II, a general (possibly linear) F0 fall,
IV. an optional final F0 upstep to the following SG

Henrichsen (2006) suggests this formalization:

i. $F_m = F_0 - \frac{m}{m_{TOTAL}} \left(F_0 - F_{m_{TOTAL}}\right)$

ii. $F'_{m,u'} = F_m + UP_{m,u'}$

iii. $F_{m,u',u} = F'_{m,u'} - \frac{u-1}{1-u} \left(F'_{m,u'} - F''_{m,u'}\right)$

iv. $F''_{m,u'} = F_{m+1} - UP_{m+1,u'}$

F_m is the fundamental frequency for the (full vowel of the) initial syllable of the *m*th SG; *u'* is the number of unstressed syllables in the *m*th SG; $F'_{m,u'}(F''_{m,u'})$ is F0 for the first (last) unstressed syllable in the *m*th SG; defined for *u'*>1 (*u'*>2); $F_{m,u',u}$ is F0 for the last unstressed syllable of the *m*th SG; defined for *u'*>3 and *u*>1. The arbitrary constants F_0, m_{TOTAL} and $F_{m_{TOTAL}}$ are all associated with linguistics properties; F_0 and $F_{m_{TOTAL}}$ are the upper and lower bound of the speaker's normal F0 range (possibly, but not necessarily a function of the sentence length too; Grønnum is not very specific here); m_{TOTAL} is the total number of SGs. The upstep function *UP* is introduced in the full papers.

3. Data-driven prosody assignment

NIP prosody assignment is based on pattern matching in a background corpus of read-aloud texts. The corpus does not include the actual sound files, but selected annotation tiers only (using the Danish PAROLE corpus, Henrichsen 2007). One speech second is thus represented by 10 8-bit numbers or so, as opposed to the 48,000 16-bit sound samples typical of UnitSyn - a data reduction of four orders of magnitude.

What data types are necessary and sufficient for reliable pattern matching? Based on pilot experiments, we settled on tiers A1 (acoustic) and L1-L5 (linguistic).

A1. Fundamental frequency (logarithmic measures)

L1. Orthographic form (dictionary approved)
L2. Phonetic form
L3. Stress pattern (stressed=2, 2ndary=1, unstr.=0)
L4. Part-of-Speech (PAROLE-style tags)
L5. Word freq. (in a 28M corpus of balanced texts)

For Danish, L1 and L5 together provide almost 100% lexical disambiguation. L2 and L3, in contrast, may vary considerably with the syntactic and semantic context. L5 was included experimentally, assuming that high-frequency tokens are more likely to appear de-stressed or time condensed than low-frequency words, grouping words otherwise unrelated in L1-L4.

3.1 The NIP algorithm presented by an example

Consider an input string *I* "du vil gerne op til slottet" (*you'd like to get (up) to the Castle*). *I* is analysed (automatically) in the dimensions L2-L5.

$L1_I$	du	vil	gerne	op	til	slottet
$L2_I$	[du]	[ve]	[gáRn0]	[Cb]	[te]	[slCd-D]
$L3_I$	2	1	2-0	2	1	2-0
$L4_I$	PRO	AUX	ADJ	ADJ	PREP	$CN_{SG,DEF}$
$L5_I$	-7.6	-5.7	-8.0	-6.3	-4.1	-11.6

Using $L1_I$-$L5_I$ as a search expression, a matching utterance *U* is identified in the NIP database:

$L1_U$	han	har	ikke	noget	imod	indvandrere
$L2_U$	[han]	[hA]	[eg0]	[nc0D]	[imoD?]	[envAndCC]
$L3_U$	1	**1**	**2-0**	**2**-0	0-**1**	**2**-1-0-0
$L4_U$	**PRO**	**AUX**	**ADJ**	PRO	**PREP**	**CN**$_{PL,-DEF}$
$L5_U$	**-5.1**	**-4.5**	-4.6	**-6.6**	-8.3	**-9.9**

Observe that, in the sound related tiers L1 and L2, *I* and *U* are unrelated; tiers L3-L5, however, show a distinct similarity (values shown in **bold**).

3.2 The NIP algorithm, summarized

Quantifying over input windows and database windows (both up to 7-place), a prosodic envelope is distilled by superimposing all envelope contributions ('envelope' = one F0 data point for each syllable) weighted by the corresponding GP^7 value. The resulting envelope is normalized wrt. permitted F0 range, duration, etc.

GP Geometrical proximity

Tier values: $GP_{TIER}(w,w')$ for $TIER = L1_x .. L5_x$
Tier vectors: $GP^{VEC}(V, V') = (\sum_{i\text{€}TIER} GP_i\ (x_i\ , x_i\ ')) /5$
Windows: $GP^7(W,W') = (\sum_{n=1..7} GP^{VEC}(V_{W,n}\ ,V_{W',n}\ ))/7$

4. Experimental evidence

16 Danish test subjects graded a suite of test sentences varied systematically for length (2-8 SGs), synthesized with the DiSyn voice Gizmo (developed at DanCAST with the festival toolkit) using prosodic models m1-m5.

m1. Grønnum's model formalized as in 2
m2. Model of DiSyn voice Carsten (www.mikrov.dk)
m3. Model of UniSyn voice Sara (www.pdc.dk)
m4. NIP based model as presented in this paper
m5. Human read-aloud version re-synthesized

The test subjects were asked to evaluate the five instancies of each sentence for naturalness: "Order the versions from best to worst" and "Grade each version as excellent/good/mediocre/bad"

As expected, all subjects preferred m5 over all other models, showing the test set-up to be reliable. Excluding m5 from the test set, these patterns emerged:

{m3,m4} were preferred over {m1,m2} by all 16 subjects, suggesting that current theory-driven models of Danish prosody are inferior to data-driven models.

13 subjects had m1>m2, suggesting our formalization of Grønnum's model to be superior to the one used in Carsten (the leading commercial DiSyn based synthetic voice for Danish).

9 subjects had m4>m3 (m4 being preferred for sentences containing several infrequent content words), suggesting that NIP-based prosodic models may offer attractive alternatives to the full-blown UnitSys system.

5. Conclusion

We do not claim NIP-driven diphone synthesis to be superior to Unit Selection *as such*. More reference data should still provide better synthesis everything else being equal. However, based on our experiments we suggest that the standard claim of huge sound databases as *necessary* remedies to the failing prosodic naturalness of diphone synthesis be reconsidered.

References

Aylett, M. P. & J. Yamagishi (2008) Combining Statistical Parametric Speech Synthesis and Unit-Selection for Automatic Voice Cloning; LangTech-2008, Rome.

Grønnum, N. (1998). Intonation in Danish. In D.Hirst et al (eds) Intonation Systems. Cambridge Univ. Press.

Guner, E.; Cenk Demiroglu (2011) A Small-footprint Hybrid Statistical and Unit Selection TTS Synthesis System for Turkish; Computer and Information Sciences vol. II; Springer

Henrichsen, P.J. (2006) Danish Prosody, Formalized. In J.Toivanen et al (2006)

Henrichsen, P.J. (2007) The Danish PAROLE Corpus - a Merge of Speech and Writing; in J. Toivanen et al (2007)

Oparin, I.; V.Kiselev; A.Talanov (2008) Large Scale Russian Hybrid Unit Selection TTS. SLTC-08. Stockholm.

Toivanen, J.; P. J. Henrichsen (eds) (2006/2007) Current Trends in Research on Spoken Language in the Nordic Countries (vol 1/2). Oulu Univ. Press.

Human evaluation of extration based summaries

Marcus Johansson, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
marjo581@student.liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

We present an evaluation of an extraction based summarizer based on human assessments of the summaries. In the experiment humans read the various summaries and answered questions on the content of the text and filled in a questionnaire with subjective assessments. The it took to read a summary was also measured. The texts were taken from the readability tests from a national test on knowledge and ability to be engaged in university studies (Sw. Högskoleprovet). Our results show that summaries are faster to read, but miss information needed to fully answer questions related to the text and also that human readers consider them harder to read than the original texts.

1. Introduction

Most evaluations of extraction based text summarizations are based on comparisons to gold standards, e.g. (Over et al., 2007; Pitler et al., 2010; Smith and Jönsson, 2011a). Such evaluations are rather straightforward to conduct and are important to assess the performance of a summarizer. The performance of the summarizer is then mainly assessed based on n-gram statistics between the gold standard and the summarization produced by the summarizer. However, such evaluations, termed intrinsic (Hassel, 2004), do not consider how readable a summary is or how much information that is conveyed from the original document.

Extrinsic evaluations, where the usability of a summary is evaluated are less common. Morris et al. (1992), however, present an evaluation of summaries where they have subjects do the American Graduate Management Aptitude test, similar to the Swedish Högskoleprovet based on summaries of varying length, a human produced abstract and no text, i.e. they have to guess when answering the questions. They did found that human produced abstracts were best, but their results were not significant. Mani et al. (1999) found that summaries comprising 17% of the original text were as good as the original text to predict if the information is relevant for a certain subject.

In this paper we present results from an extrinsic evaluation of an extraction based summarizer based on experiments with humans answering test questions after reading the original text, summaries and guessing.

2. Method

We recruited 60 students, mean age 22.6 years, 22 women and 38 men, all at Linköping University. The test was similar to the test by Morris et al. (1992). We used the reading comprehension test from the National test for high school studies (Sw. Högskoleprovet) from 2011 as we assumed that none of them had previous experience with that test set, which they did not have. The experiment used four different test sets. One extra test set was used as training set. The four test sets were summarized using the extraction based summarizer CogSum (Smith and Jönsson, 2011b) to 30% of the original text length. We also kept the original text for comparisons. The original texts were between 928-1108 words and the summaries between 410-308 words.

Before the test the subjects answered a questionnaire comprising background data such as age, sex, if they have done the test before and esteemed reading ability. They then practiced on a text that was not used in the actual test and after that they did the actual test under three conditions: 1) reading a 30% summarization, 2) reading the original text and, 3) to answer the questions without any text at all, i.e. they had to guess. These three conditions were handed to the subjects in a different and balanced order between the subjects.

For the text conditions the subjects first read the tests' pre-defined questions, then they were handled the text which included answers to the questions and finally the questions were handed back to them and they were asked to answer them. We measured the time it took for the subjects to carry out each sub task, read questions, read text and answer the questions. After the test the subjects had to answer a second questionnaire with Likert scale items (1-7) on their attitudes towards the text such as how easy it was to read, if all relevant information was in the text, if it took long to read and understand etc. We also measured the number of correct answers to the questions.

3. Results

Table 1 shows the number of correct answers on the questions in the test *Högskoleprovet*. The data were analyzed using a within-group ANOVA test which gave the result $F(1.86, 109.77) = 30.735, p < .01, \eta^2 = .34$, Huynh-Feldt corrected. This was followed by a SIDAK post-hoc test to investigate differences between conditions.

Table 1: Number of correct answers and time to read for each text type.

Text type	Correct answers		Time(sec)	
	Mean	StDev	Mean	Stdev
Original text	2.62	1.04	337.6	109.38
Summary	2.2	1.03	153.9	61.34
Guessing	1.3	0.94		

Table 2: Means and standard deviations for questionnaire items for the original text and the summary text.

Item	Original		Summary	
	Mean	StDev	Mean	Stdev
I think the text gives a good conception of the subject	4.63	1.52	3.20	1.37
I experience the text as information rich	4.70	1.37	3.48	1.56
I think the text has a good flow	4.75	1.49	3.63	1.69
I experience that the text misses relevant information in order to answer the questions	3.25	1.44	4.55	1.65
I think the text was easy to comprehend	4.93	1.59	4.12	1.60
I think it took a long time to read the text	3.87	1.33	3.28	1.32
I think the text was easy to read	4.78	1.61	4.08	1.81
I think the text was exhausting to read	3.55	1.67	3.85	1.67

As expected reading the original text gives significantly more correct answers than reading the summary, $p < .05$. Both the summary and the original text give significantly more correct answers, $p < .001$. The difference between the original text and the summary was 10.5% fewer correct answers.

Table 1 also depicts the time it took to read the text. The time to read the summary, is 55% shorter than the time it takes to read the original text, a significant difference $t(59) = 17.73, p < .001$. This is true even if we calculate the time it took to read 30% of the original text, $t(59) = 9, p < .001$.

We did not find any significant difference in the time it took to answer the test questions.

Table 2 depicts the subjective scores on the items in the questionnaire for reading the original text and the summary.

There were statistically significant differences, two-tailed t-test significance level $p < .001$, between the original text and the summary for all items (all $ts > 2.5$, all $ps < .01$), except for *I think the text was exhausting to read* where no significant difference was found.

4. Discussion

We have presented results from an evaluation of extraction based summaries. Sixty subjects read texts from a national readability test and answered questions on the text. In the study we also measured reading time and the subjects answered a questionnaire with items on their perceived text quality.

When our subjects read the original text they had significant more correct answers to the questions than they had when reading a 30% summary of the text. However, compared to guessing without reading the text the subjects had significantly more correct answers after reading the summary. Furthermore, the amount of information lost in the summary compared to the original text is only 10%, and considering that 70% of the text is lost in the summary this can be considered as an acceptable loss of information, especially as the time it took to read the text was around 50% shorter reading the summary compared to reading the full text.

The importance of losing information depends, of course, on the type of text. Persons reading a news text probably accept losing 10% or even more of the text, especially if it means saving 20% of the time it takes to read. For other texts, such as texts on how to fill in authority forms or the texts we used on taking a test, we can assume that information loss is more problematic.

Overall the original texts were considered better than the summaries. They were easier to read, had a better cohesion, and contained more information. The mean values on the various Likert items for the summaries, as seen in Table 1, are often around 3.5, i.e. the arithmetic mean of the scale with a maximum of 7 and although significantly worse than the originals the difference is only about 1 point on the scale indicating that the summaries are not that bad.

One important target group for automatic text summarization is persons with reading difficulties, such as dyslectics. Conducting studies with such persons is an important future work.

5. References

Martin Hassel. 2004. Evaluation of automatic text summarization. Licentiate Thesis, 3-2, Stockholm University.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of EACL-99*.

Andrew H. Morris, George M. Kasper, and Dennis A. Adams. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43:1506–1520, Jan.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality inmulti-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 544–554.

Christian Smith and Arne Jönsson. 2011a. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*.

Christian Smith and Arne Jönsson. 2011b. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*.

Bridging the Gap between Two Different Swedish Treebanks

Richard Johansson

Språkbanken, Department of Swedish, University of Gothenburg
Box 100, SE-40530 Gothenburg
richard.johansson@gu.se

Abstract

We present two simple adaptation methods to train a dependency parser in the situation when there are multiple treebanks available, and these treebanks are annotated according to different linguistic conventions. To test the methods, we train parsers on the Talbanken and Syntag treebanks of Swedish. The results show that the methods are effective for low-to-medium training set sizes.

1. Introduction

When developing a data-driven syntactic parser, we need to fit the parameters of its statistical model on a collection of syntactically annotated sentences – a *treebank*. Generally speaking, a larger collection of examples in the training treebank will give a higher quality of the resulting parser, but the cost in time and effort of annotating training sentences is fairly high.

However, there is an abundance of theoretical models of syntax and there is no consensus on how treebanks should be annotated. For some languages, there exist multiple treebanks annotated according to different syntactic theories. The list of such languages includes Swedish, Chinese, German, and Italian. Given the high cost of treebank annotation and the importance of a proper amount of data for parser development, this situation is frustrating.

In this paper, we investigate two simple adaptation methods to bridge the gap between differing syntactic annotation styles, in order to be able to use more data for parser training. As a case study, we train dependency parsers on two Swedish treebanks.

2. Two Swedish Treebanks

As previously noted, by Nivre (2002) *inter alia*, Swedish has a venerable tradition in treebanking: there are not only one but two treebanks which must be counted among the earliest efforts of that kind. The oldest one is the Talbanken or MAMBA treebank (Einarsson, 1976), which has later been reprocessed for modern use. The original annotation is a function-tagged constituent syntax without phrase labels, but the reprocessed release includes a version converted to dependency syntax.

The second treebank is called Syntag (Järborg, 1986). Similar to Talbanken, its representation is a function-tagged constituent formalism without phrase labels. We developed a conversion to dependency trees, which was fairly straightforward since many constituents are annotated with heads.

The two treebank annotation styles have significant differences. Most prominently, the Syntag annotation is fairly semantically oriented in its treatment of function words such as prepositions and subordinating conjunctions: in Talbanken, a preposition is the head of a prepositional phrase, while in Syntag the head is the prepositional complement.

3. Training Parsers on Multiple Treebanks

We now describe the two adaptation methods to leverage multiple treebanks. In this work, we use a common graph-based parser (Carreras, 2007). In such a parser, for a given sentence x, we find the top-scoring parse $\hat{y}$ that maximizes a linear function $w \cdot f(x, y)$, where w is a weight vector produced by some learning algorithm and $f(x, y)$ a *feature representation* of the sentence x and its parse tree y. The adaptation methods presented in this work consist of modifications of the feature representation function f.

3.1 Using a Common Feature Representation

Our first adaptation method is a slightly generalized idea from domain adaptation techniques (Daumé III, 2007): let the machine learning algorithm find which properties of the two datasets are common and which are different.

In practice, this is implemented as follows. Assume that a sentence x with a parse tree y is represented as $f_1(x, y)$ if it comes from the first treebank, and $f_2(x, y)$ if from the second treebank. We then add a *common feature representation* f_c to f_1 and f_2, and embed them into a single feature space. The resulting feature vectors then become $\langle f_1(x, y), \mathbf{0}, f_c(x, y)\rangle$ for a sentence from the first treebank, and $\langle \mathbf{0}, f_2(x, y), f_c(x, y)\rangle$ for the second treebank. Using this representation, the two datasets are combined and a single model trained. The hope is then that the learning algorithm will store the information about the respective particularities in the weights for f_1 and f_2, and about the commonalities in the weights for f_c.

In this work, f_1, f_2, and f_c are identical, as in the original formulation by Daumé III (2007): all of them correspond to the feature set described by Carreras (2007). However, it is certainly imaginable that f_c could consist of specially tailored features that make generalization easier.

3.2 Using One Parser to Guide Another

The second method is inspired by work in parser combination. For instance, Nivre and McDonald (2008) combined a graph-based and a transition-based parser by adding *guide features*: features that describe the output of one parser when running the other one. The resulting parsers combine the strengths of the two different parsing approaches.

We added guide features to the parser feature representation. However, the features by Nivre and McDonald (2008) are slightly too simple since they only describe whether two words are attached or not; this will not help us if there are systematic differences between the two treebanks. Instead, we used *path* features as in semantic role labeling. For a possible head h and dependent d, sibling s and grandchild g, we extracted the following features:

- POS(h)+POS(d)+Path($h \to d$)
- POS(h)+POS(s)+Path($h \to s$)
- POS(h)+POS(d)+POS(s)+Path($h \to s$)
- POS(h)+POS(g)+Path($h \to g$)
- POS(h)+POS(d)+POS(g)+Path($h \to s$)

4. Experiments

We first trained standard parsers on Talbanken and Syntag. These parsers perform fairly well, although the accuracy of the Syntag parser is lower due to the smaller size and semantically oriented annotation of its training set. However, when we run the Talbanken parser on the Syntag test set or vice versa, the parsers perform very poorly. Combining the two training sets results in a parser performing poorly on both sets. All attachment accuracies are given in Table 1.

Training set	Acc. ST	Acc. TB
Syntag	83.0	52.4
Talbanken	50.3	88.3
Combined	59.9	83.6

Table 1: Performance figures of the baseline parsers.

The effects of the two adaptation methods are illustrated in Figures 1 and 2. In Figure 1, we show the learning curves of the parsers trained on the two different treebanks. We see that the adaptation methods lead to learning curves that rise faster, making more efficient use of the data for small treebanks. However, for both treebanks the effect is insignificant when the treebanks are larger.

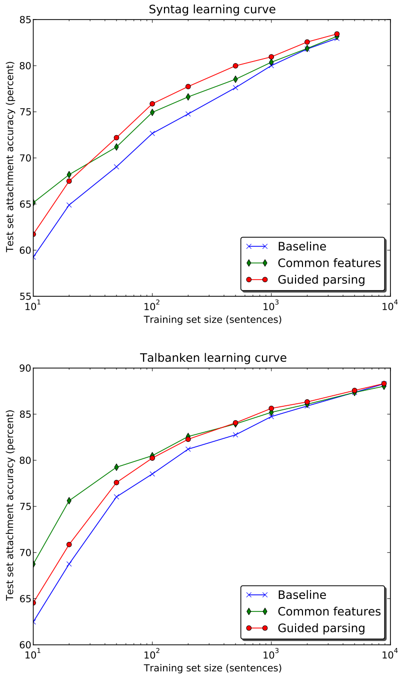


Figure 1: Attachment accuracy by training set size.

Figure 2 shows the error reduction percentages of the adaptation methods as a function of the training set size. This plot nicely illustrates the different properties of the two methods: using a common feature representation is more effective when the training sets are extremely small, while using guided features works best for mid-sized sets.

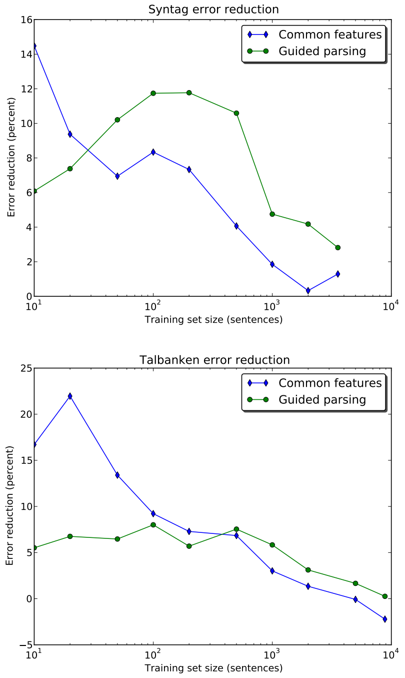


Figure 2: Error reduction of the adaptation methods as a function of training set size.

5. Discussion

We have studied two very simple cross-framework adaptation methods that allow us to make use of multiple treebanks when training parsers. Both methods give large improvements for small training set sizes, while the effect gradually diminishes as the training set size increases. While this work used an unlabeled graph-based dependency parser, the methods presented here generalize naturally to other parsing methods including transition-based parsing as well as discriminative constituent parsing.

6. References

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings the CoNLL Shared Task*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Jan Einarsson. 1976. Talbankens skriftspråkskonkordans. Department of Scandinavian Languages, Lund Univ.

Jerker Järborg. 1986. Manual för syntaggning. Department of Linguistic Computation, Univ. of Gothenburg.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL*.

Joakim Nivre. 2002. What kinds of trees grow in Swedish soil? A comparison of four annotation schemes for Swedish. In *Proceedings of TLT*.

SCXML for Building Conversational Agents in the Dialog Web Lab

David Junger, Torbjörn Lager and Johan Roxendal

Department of Philosophy, Linguistics and
Theory of Science, University of Gothenburg.
Department of Swedish, University of Gothenburg.
tffy@free.fr, lager@ling.gu.se, johan.roxendal@gu.se

1. Introduction

The W3C has selected Harel Statecharts, under the name of State Chart XML (SCXML), as the basis for future standards in the area of (multimodal) dialog systems (Barnett et al. 2012). In an effort to educate people about SCXML we are building a web-based development environment where the dialogs of embodied, spoken conversational agents can be managed and controlled using SCXML, in a playful and interesting manner.

2. SCXML = STATE CHART XML

SCXML can be described as an attempt to render Harel statecharts (Harel 1987) in XML. Harel developed his statecharts as a tool for specifying reactive systems in great detail. In its simplest form, a statechart is just a finite state machine (FSM), where state transitions are triggered by events appearing in a global event queue.

Just like ordinary FSMs, statecharts have a graphical notation. Figure 1 depicts a very simple example:

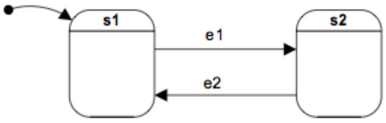


Figure 1: A simple statechart.

Any statechart can be translated into a document written in the linear XML-based syntax of SCXML. Here, for example, is the SCXML document capturing the statechart in Figure 1:

```
<scxml>
  <state id="s1">
    <transition event="e1" target="s2"/>
  </state>
  <state id="s2">
    <transition event="e2" target="s1"/>
  </state>
</scxml>
```

The document can be executed by an SCXML conforming processor, providing for just a small step from a specification into a running application.

Harel (1987) also introduced a number of (at the time) novel extensions to FSMs, which are also present in SCXML. Most importantly, the ability to specify hierarchical state machines as well as machines running in parallel takes care of some of the problems of ordinary FSMs – in particular the notorious state explosion problem. Furthermore, a complex state may contain a *history state*, serving as a memory of which substate S the complex state was in, the last time it was left for another state. Transition to the history state implies a transition to S. In addition, the SCXML standard also provides authors with means for accessing external web-API:s to for example databases.

The expressivity of SCXML makes it possible not only to specify large and very complex dialog managers in the style of FSMs, but also to implement more sophisticated dialog management schemes such as the Information-State Update approach (Kronlid & Lager 2007).

The work on SCXML is almost completed and several implementations exist. At the time of writing, the most conforming implementation, written by the third author, is called PySCXML and is implemented in Python. We are also aiming at an implementation in JavaScript, and a first version exists, written by the first author in collaboration with the third. The latter implementation is used in the Dialog Web Lab.

3. Speech recognition in Google Chrome

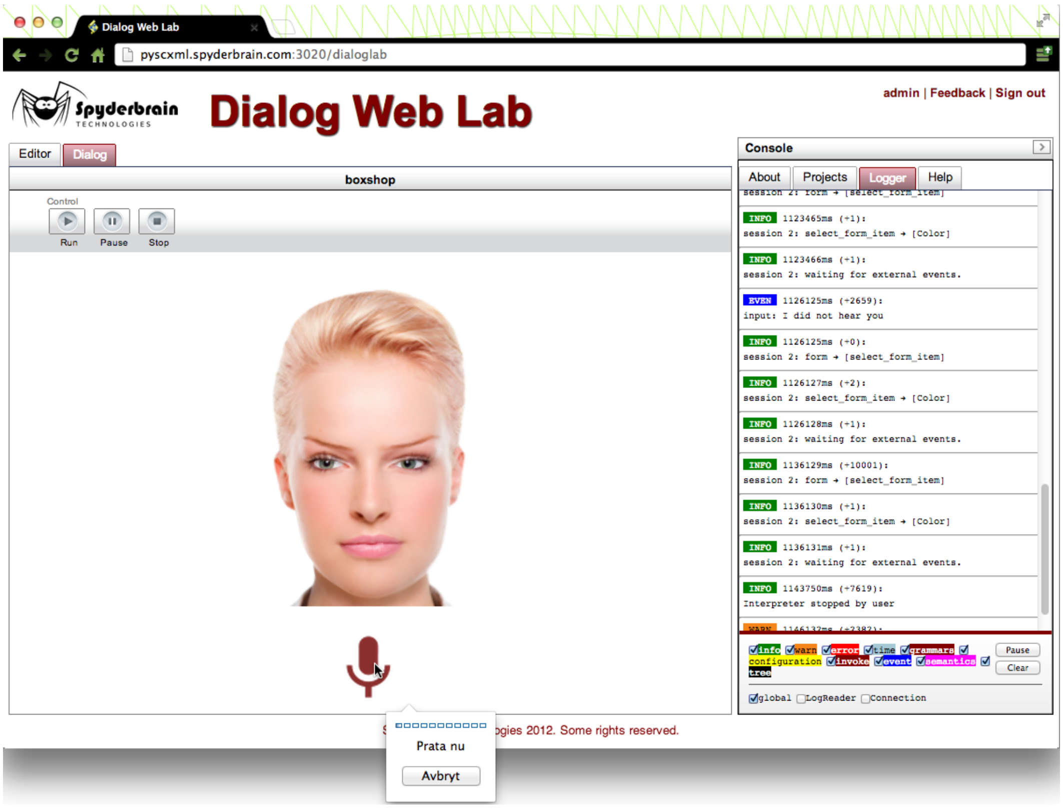
Recent versions of the Google Chrome browser offer web developers additional markup for creating <input> fields into which users can either write something in the usual way, or, by clicking a small microphone icon that is part of the widget, speak something that is transcribed into text. We have styled the widget by means of CSS and use it in the Dialog Web Lab in the way shown in Figure 2.

4. Speech synthesis and animated faces

SitePal (www.sitepal.com) is a commercial cloud platform for building speaking avatars, developed by Oddcast, that allows users to deploy “virtual employees" on websites that can welcome visitors, guide them around the site and answer questions. The avatars move their lips in synch with the speech, are capable of some non-verbal behavior, and can express emotions through facial expressions. The speech and bodily behavior of avatars can be controlled from JavaScript and thus also from our SCXML implementation.

5. The Dialog Web Lab

The idea behind the Dialog Web Lab is simply to allow developers of conversational agents access to an application



```
session 2: form → [select_form_item]
INFO 1123465ms (+1):
session 2: select_form_item → [Color]
INFO 1123466ms (+1):
session 2: waiting for external events.
EVEN 1126125ms (+2659):
input: I did not hear you
INFO 1126125ms (+0):
session 2: form → [select_form_item]
INFO 1126127ms (+2):
session 2: select_form_item → [Color]
INFO 1126128ms (+1):
session 2: waiting for external events.
INFO 1136129ms (+10001):
session 2: form → [select_form_item]
INFO 1136130ms (+1):
session 2: select_form_item → [Color]
INFO 1136131ms (+1):
session 2: waiting for external events.
INFO 1143750ms (+7619):
Interpreter stopped by user
WARN 1146132ms (+2382):
```

☑info ☑warn ☑error ☑time ☑grammars ☑configuration ☑invoke ☑event ☑semantics ☑tree

☑global □LogReader □Connection

Figure 2: The current version of the Dialog Web Lab.

where they can explore the use of SCXML as a tool for dialog management.

In the current version of the Dialog Web Lab, depicted in Figure 2, the user has authored an SCXML document, is now executing it, and has just clicked the microphone. The user may now speak to the agent which, controlled by the SCXML process, will respond using speech and possibly non-verbal behavior. The Logger is activated, allowing the user to trace the execution in great detail.

6. Future work

We see several ways in which the Dialog Web Lab could be improved and extended:

- The present version of the Dialog Web Lab only works in recent versions of Google Chrome. However, The HTML Speech Incubator Group, in collaboration with the Voice Browser Working Group, has begun to specify a standard that integrates speech technology in HTML5, in the form of both markup and a JavaScript API. The speech recognition implemented in Google Chrome should, we believe, be seen as a preview of what to expect. At this point in time, the work on the standard has only just started, but once it is finished, and widely implemented, this will likely mean that the Dialog Web Lab can take advantage of it.

- In a future version of the Dialog Web Lab we plan to allow a user to control more than one conversational agent at the same time. This might be an interesting way to exercise the parallelism available in SCXML, and an interesting way to experiment with multi-party dialog involving conversational agents.

- As it turns out, a large fragment of VoiceXML can very easily be compiled into SCXML. We plan to allow authors to mix SCXML and VoiceXML in one document, compile it into SCXML and then run it. We believe that this will make possible very succinct expressions of dialog management strategies.

- The statechart formalism is a graphical language. So one idea would be to allow authors to draw statecharts on a canvas, compile them into SCXML, and then run them. Unfortunately, actually building such a graphical editor that works in a browser is not easy.

7. References

Barnett, Jim (Ed.) (2012) State Chart XML (SCXML): State Machine Notation for Control Abstraction, W3C Working Draft 16 February 2012. <http://www.w3.org/TR/scxml/>

Harel, David (1987) Statecharts: A Visual Formalism for Complex Systems, In: *Science of Computer Programming* 8, North-Holland.

Kronlid, Fredrik and Lager, Torbjörn (2007). Implementing the Information-State Update Approach to Dialogue Management in a Slightly Extended SCXML. In Ron Artstein and Laure Vieu (Eds.) *Proceedings of the 11th International Workshop on the Semantics and Pragmatics of Dialogue* (DECALOG), Trento, Italy, s. 99-106.

Generating Pronunciations of Northern Saami Words by Two-Level Rules

Pekka Kauppinen
University of Helsinki
Helsinki, Finland
pekka.kauppinen@helsinki.fi

1. Introduction

In my research (Kauppinen, 2012), I have examined the possibility of using a finite-state transducer compiled from a two-level grammar to generate pronunciations of written Northern Saami word forms.

2. Northern Saami

The Northern Saami language distinguishes three phonemic consonant quantities: short, long and overlong, from here on referred to as Q1, Q2, and Q3, respectively. The quantity of a consonant or a consonant cluster is usually reflected in the quantity of the preceding stressed vowel or diphthong (Bals et al., 2007). Consonant segments preceded by a stressed vowel are subject to a process known as consonant gradation – in most cases this means that Q3 consonant segments alternate with Q2 segments and Q2 segments with Q1 segments in certain inflected forms:

juolgi 'foot' (Q3)	: *juolggit* 'feet' (Q2)
beassi 'nest' (Q2)	: *beasis* 'in the nest' (Q1)
giehta 'hand' (Q2)	: *gieđat* 'hands' (Q1)

The current orthographic standard for Northern Saami was approved by the Nordic Saami Council in 1979 and is now used in all the Nordic states where the language is spoken. It was designed not only to be both simple and largely phonemic but also to unify the different varieties of the language by "smoothing out details of actual pronunciation that are not shared by all dialects of North Saami" (Bals et al., 2007), which allows each speaker to pronounce the written language according to their dialect of preference.

On the other hand, a simpler orthography with as few special symbols as possible has also led to a situation where certain qualitative and quantitative distinctions are not marked at all. Consider the following examples:

beassi	'tree bark'(Q3)	[pĕæ̆sː(s)i·]
beassi	'tree bark's' (Q2)	[peæsːi·]

In general, the distinction between Q3 and Q2 geminates, although audible in speech, is not marked by the current orthography. Other features that are not marked include vowel length (with the exception of *a* [ɑ] and *á* [ɑː]) and the distinction between the consonant cluster /lj/ as in *olju* [olºju·] 'oil' and the more common long palatal lateral /ʎ:/ as in *vielja* [ʋĭeʎːɑ] 'brother', both spelled *lj*.

Apart from the aforementioned shortcomings, however, the relationship between spelling and pronunciation can be expressed in terms of rules – as done by Sammallahti (1981), for instance – which makes it fair to say that the current orthographic convention does maintain a certain degree of regularity in relation to pronunciation.

3. Resources

I have chosen the well documented Eastern Eanodat dialect as the foundation of the pronunciations in my research. My primary sources were Pekka Sammallahti's description of the dialect's phonetics and phonology (1977) as well as his Northern Saami textbook (1981).

I have mainly used HFST (*Helsinki Finite-State Transducer Technology*), and Måns Huldén's Foma toolkits to create finite-state transducers from lexica (*lexc* files) and two-level grammars.

4. Lexica

The first step in generating pronunciations of Northern Saami words is to supplement the plain written forms with the information that might be missing from them due to the current orthographic convention. These supplemented forms are then processed by another finite-state transducer that executes the actual grapheme-to-sound conversion. The latter FST is compiled from the two-level grammar containing the rules that describe the general correspondence between Northern Saami orthography and the Eastern Eanodat pronunciation.

The problem of recognizing Q3 geminates as well as distinguishing *lj* /lj/ from *lj* /ʎ:/ can be largely solved by building a lexicon of word forms containing such sequences and compiling it into an FST. However, morphological analysis and syntactic disambiguation may be needed to determine which interpretation of an ambiguous case such as *guossi* – which could represent either *guossi* 'guest' (Q3) or *guossi* 'guest's' (Q2) – is possible or acceptable in a certain context.

The same method could be used to distinguish long monophthongs from their invariably short counterparts, although this might not be the best solution possible, since most long vowels occur as allophones of diphthongs in certain easily recognizable morphophonological environments and are also much more common than Q3 geminates. I decided not to delve deeper into this problem in my thesis as I felt it would deserve a paper of its own.

The pronunciation of some recent loanwords may also differ depending on dialect or speaker even though there may be no difference in spelling. A straightforward way of resolving this problem would be by building a dialect-specific lexicon of these cases and compile it into an FST.

In my research, I have used a FST compiled from a small lexicon of the 40 common forms containing Q3 geminates and long vowels in order to supplement the word forms entered as input. These 40 word forms were taken from Giellatekno's list of 10 000 most frequent word forms which is freely available at the project's website (http://giellatekno.uit.no/). Naturally, a lexicon of this size would by no means be sufficient for a finalized application.

5. The TWOL grammar

In a two-level grammar, the correspondence between a symbol representing the deep form and the one representing its surface form is expressed as a symbol pair, in this case [grapheme]:[sound] (Beesley et al., 2003). Two-level rules are constraint rules that operate at two levels only without intermediate phases, the set of rules functioning as a filter that only allows the desired sequences of grapheme-sound pairs to pass (Koskenniemi, 1983).

The two-level grammar I used the generate the pronunciations consisted of 45 rules and was written in limited spare time over the course of one month, with some minor improvements made afterwards. Sammallahti's Northern Saami textbook not only included a detailed description of the relationship between spelling and the Eastern Eanodat pronunciation but also provided 144 written forms with phonetic transcriptions, which I used as reference material when writing my two-level grammar.

At its simplest, the task of writing the two-level rules was merely about rewriting Sammallahti's rules by using the TWOL formalism. For example, one rule states that the grapheme *a* (symbol a) is pronounced like as a long open vowel [ɑː] (symbol ā) in an even syllable if the vowel of the preceding odd syllable is short and there is a simple consonant between the two vowels. Written as a TWOL rule, it looks something like this:

```
a:ā <=> OddSyll :ShortV :Cons _ ;
```

Here the element called OddSyll is equivalent to a sequence of even number of syllables. The context :ShortV stands for [:a | :o | :e | :u | :i], and Cons is the set of consonant symbols in the alphabet.

6. Evaluation

Two-level rules have been used before for grapheme–sound conversion for e.g. Welsh and Turkish. However, no method has been described for Northern Saami, and it appears that two-level rules have not been used for this purpose either. In any case, an FST that converts a ortographic forms into phonetic forms and vice versa has uses in the field of speech synthesis as well speech recognition technology, none of which exists for Northern Saami at the present time.

I combined the two FSTs into a single transducer, which I tested with a Northern Saami text sample of 216 words. The results are shown below (Table 1).

Recall	~ 99,5 %
Precision	~ 97,7%

Table 1. Preliminary results.

The FST yielded a total of 215 pronunciations, of which 210 were correct. Four of the remaining word forms contained an incorrect consonant or vowel quantity due to the small size of the lexicon used. One word form, *leatge,* is pronounced differently from what output suggested because the *-ge* at the end is a clitic – again, morphological analysis would have been helpful here. A single word form was not generated at all due to an error or a contradiction in the two-level grammar. Indeed, while a large TWOL grammar is easier to maintain and expand than e.g. a sizeable set of rewrite rules (since the former are applied in parallel and are less likely to affect each other), the deterministic nature of the method calls for unambiguous rules that may be cumbersome to formulate.

I restricted my research to stand-alone word forms instead of phrases or sentences, and thus neglected several lenition and assimilation processes triggered by surrounding words – an example would be as the lenition of a word-initial *d* /t/ in certain words when preceded by a vowel or a glide (Sammallahti, 1977). This, along with the problems mentioned earlier, should be fixed and given more attention if the two-level grammar described here is to be used in practical applications.

7. Conclusion

In short, my research concluded that two-level rules are a valid method of generating pronunciations of Northern Saami words. However, they are not necessarily superior to any other method available. The problem of supplementing word forms with the information not indicated in spelling remains regardless of the method used for grapheme-to-sound conversion.

8. Reference

Bals, B., Odden, D. and Rice, C. (2007). The Phonology of Gradation in North Saami. From the series *Topics in North Saami Phonology*. Universitet i Tromsø.

Beesley, K. and Karttunen, L. (2003). Two-Level Rule Compiler. Document available at http://www.stanford.edu/~laurik/.book2software/twolc.pdf. Retrieved the 5th of October, 2012.

Kauppinen, P. (2012). Pohjoissaamen ääntämysasujen tuottaminen kaksitasosäännöillä. BA thesis. Department of Modern Languages. University of Helsinki.

Koskenniemi, K. (1983). Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics. University of Helsinki.

Sammallahti, P. (1977). Norjansaamen Itä-Enontekiön murteen äänneoppi. Helsinki.

Sammallahti, P. (1981): Saamen peruskurssi 1. Kokeilumoniste. Kouluhallitus. Helsinki.

Automatic Text Simplification via Synonym Replacement

Robin Keskisärkkä, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
robin.keskisarkka@liu.se, arnjo@ida.liu.se

Abstract

Automatic lexical simplification via synonym replacement in Swedish was investigated. Three different methods for choosing alternative synonyms were evaluated: (1) based on word frequency, (2) based on word length, and (3) based on level of synonymy. These three strategies were evaluated in terms of standardized readability metrics for Swedish, average word length, and proportion of long words, and in relation to the ratio of type A (severe) errors in relation to replacements.

1. Introduction

In this paper we present results from an investigation on whether a text can be successfully simplified using synonym replacement on the level of one-to-one word replacements. Theoretically, synonym replacements can affect established readability metrics in Swedish in different ways. The correlation between word length and text difficulty indicates that lexical simplification is likely to result in decreased word length overall, and a decrease in number of long words. Also, if words are replaced with simpler synonyms we can expect a smaller variation in terms of unique words, since multiple nuanced words may be replaced by the same word.

Studies within lexical simplification have historically investigated the properties of English mainly, and almost all rely in some way on the use of WordNet (Carroll et al., 1998; Lal and Rüger, 2002; Carroll et al., 1999). For Swedish there is no WordNet or system of similar magnitude or versatility. A few studies have used lexical simplification as a means of simplifying texts to improve automatic text summarization (Blake et al., 2007), and some have applied some type of lexical simplification coupled with syntactic simplification, but studies that focus on lexical simplification in its own right are rare. The studies that do exist tend to view lexical simplification as a simple task in which words are replaced with simpler synonyms, defining a *simpler* word as one that is more common than the original.

Words with identical meaning in all contexts are rare and any tool that replaces words automatically is therefore likely to affect the content of the text. This does not mean that automatic lexical simplification could not be useful, e.g. individuals with limited knowledge of economy may profit little by the distinction between the terms *income*, *salary*, *profit*, and *revenue*.

2. Method

We use the freely available SynLex in which level of synonymy between words is represented in the interval 3.0–5.0, where higher values indicate a greater level of synonymy between words. The lexicon was constructed by allowing Internet users of the Lexin translation service to rate the level of synonymy between Swedish words on a scale from one to five (Kann and Rosell, 2005). SynLex was combined with Parole's frequency list of the 100,000 most common Swedish words by summarizing the the frequencies of the different inflections of the words in the synonym dictionary. The final file contained synonym pairs in lemma form, level of synonymy between the words, and word frequency count for each word. The original synonym file contained a total of 37,969 synonym pairs. When adding frequency and excluding words with a word frequency of zero 23,836 pairs remained.

Readability was evaluated using LIX, OVIX, average word length, and proportion of long words. The texts were checked for errors manually, using a predefined manual. Inter-rater reliability, between two raters, was 91.3%.

Errors were clustered into two separate categories: *Type A errors* include replacements which change the semantic meaning of the sentence, introduce non-words, introduce co-reference errors within the sentence, or introduces a different word class (e.g. replaces a noun with an adjective). *Type B errors* consist of misspelled words, article or modifier errors, and erroneously inflected words.

Text were chosen from four different genres: newspaper articles from *Dagens nyheter* (DN), informative texts from Försakringskassan's homepage (FOKASS), articles from *Forskning och framsteg* (FOF), and academic text excerpts (ACADEMIC). Every genre consisted of four different documents which were of roughly the same size. The average text contained 54 sentences with an average of 19 words per sentence. In the experiments synonym replacement was performed on the texts using a one-to-one matching between all words in the original text and the available synonyms. A filter was used which allowed only open word classes to be replaced, i.e. replacements were only performed on words belonging to the word classes nouns, verbs, adjectives, and adverbs.

In the first two experiments the three conditions word frequency, word length, and level of synonymy are used to choose the best replacement alternative. The first strategy compares word frequencies and performs substitutions only if the alternative word's frequency is higher than that of the original, if more than one word meets this criteria the one with the highest word frequency is chosen. Similarly, word length replaces a word only if the alternative word is shorter, if more than one word meets the criteria the shortest one is chosen. The third strategy replaces every word with the synonym that has the highest level of synonymy. In experiment two the inflection handler is introduced. The inflection handler allows synonym replacement to be performed based on lemmas, which increases the number of potential replacements. The inflection handler also functions as an extra filter for the replacements since only words that have an inflection form corresponding to that of the word being replaced are considered as alternatives. In the third experiment thresholds are introduced for the different strategies. The thresholds are increased incrementally and the errors are evaluated for every new threshold. Finally, in the fourth experiment word frequency and level of synonymy are combined and used with predefined thresholds.

3. Results

The results from experiment one showed that for all genres the replacement based on word frequency resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was 0.52. Replacement based on word length also resulted in an improvement in terms of readability for every genre in all readability metrics. The average error ratio was 0.59. For all genres the replacement based on level of synonymy affected the readability metrics negatively for all metrics, except for the OVIX-value. The average error ratio was 0.50.

The results from experiment two showed that for all genres the replacement based on word frequency resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was highest for FOF, 0.37, and lowest for FOKASS, 0.31. The average error ratio was 0.34. For all genres the replacement based on word length resulted in an improvement in terms of readability for every genre in all readability metrics. The error ratio was highest for FOKASS, 0.47, and lowest for FOF, 0.37. The average error ratio was 0.42. For all genres the replacement based on level of synonymy affected the readability metrics negatively for all genres in all readability metrics, accept for FOF where the OVIX-value decreased slightly. The error ratio was most highest for FOF, 0.46, and lowest for DN, 0.40. The average error ratio was 0.44.

Experiment three revealed no clear relationship between threshold and error ratio for any of the three replacements strategies. For some texts the error ratio decreased as the the threshold increased, while for others the opposite was true, and the ratio of errors remained relatively constant.

In experiment four we combined word frequency and level of synonymy. The frequency threshold was set to 2.0, meaning that only words with a frequency count of two times that of the original word were considered possible substitutions. The threshold for level of synonymy was set to 4.0. The experiment was run twice, prioritizing either word frequency (PrioFreq) or level of synonymy (PrioLevel) when more than one word qualified as an alternative. The same words are replaced in both cases, but the word chosen as the substitution may differ. In both runs the average error ratio was 0.27. PrioLevel performed significantly better than the frequency strategy in experiment two in terms of error ratio when looking at all texts and genre was not considered. Also, both PrioLevel and PrioFreq performed significantly better than the frequency strategy alone when looking at the genre DN specifically.

4. Discussion

The overall error ratio of replacing synonyms based on frequency is not significantly affected by the introduction of relative threshold frequencies for alternative words. As the frequency threshold is increased the new words are more likely to be familiar to the reader, but this does not significantly increase the likelihood that the replacement is a correct synonym in the particular context. Word length as a strategy for synonym replacement improves the text in terms of the readability metrics, but it is not clear whether it contributes to the actual readability of the text. Also, the combination of frequency and level of synonymy slightly improves the error ratio compared to frequency alone.

The study shows that the common view of automatic lexical simplification as a task of simply replacing words with more common synonyms results in a lot of erroneous replacements. The error ratio does not critically depend on level of synonymy, rather the overall error ratio remains roughly the same even when using words with the highest level of synonymy. The high error ratios at this level confirm that the concept of synonyms is highly dependent on the context. Handling word collocations and word disambiguation could greatly improve both the quality of the modified texts and substantially decrease the error ratio, but this is by no means a trivial task.

A simplified text can potentially be useful, even if it contains some errors, especially if the original text is too difficult to comprehend for the unassisted reader. It would be interesting to study the sensitivity of readers to typical erroneous replacements, and the effects simplification has on comprehension. Future studies should also aim at replacing only those words which are regarded difficult to a particular reader, rather than trying to simplify all words.

5. References

Catherine Blake, Julia Kampov, Andreas K Orphanides, David West, and Cory Lown. 2007. Unc-ch at duc 2007: Query expansion, lexical simplification and sentence selection strategies for multi-document summarization. *Proceedings of Document Understanding Conference (DUC) Workshop 2007.*

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, volume 1, pages 7–10. Citeseer.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270.

Viggo Kann and Magnus Rosell. 2005. Free construction of a free swedishdictionary of synonyms. In *NoDaLiDa 2005*, pages 1–6. QC 20100806.

Patha Lal and Stefan Rüger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL.*

Falling Ill and the *Administration of Medication*... A Study on the Nature of Medical Frames

Dimitrios Kokkinakis
Språkbanken, Department of Swedish
University of Gothenburg, Sweden, Box 200, SE 405 30 Göteborg
dimitrios.kokkinakis@svenska.gu.se

Abstract

Natural language understanding (NLU), that is enabling computers to derive meaning from natural language input, has been a long-term goal for Natural Language Processing. NLU systems require a semantic theory to guide the comprehension of the text and at the same time a suitable framework for representing lexical knowledge, preferably linked to domain ontologies. In this context, such a framework could play a vital role for alleviating the extraction of semantic relations and events. This paper suggest that FrameNet is such a framework that can facilitate the development of text understanding and as such can be used as a backbone to NLU systems. We present initial experiments of using domain-specific FrameNet extensions for the automated analysis of meaning in medical texts.

1. Introduction

Event-based, or event-template information extraction initiated by and explored in the MUC-4 extraction task (Rau *et al.*, 1992). Since then, extraction and labeling of events has also attracted attention in various shared tasks (Ruppenhofer *et al.*, 2010). Lately, algorithms are developed that try to instead learn template structures automatically from raw text; Chambers & Jurafsky (2011). Mining complex relations and events has also gained a growing attention in specialized domains, such as biomedicine; Ananiadou *et al.* (2010) and for several reasons. A publication volume that increases at an exponential level, the availability of mature NLP tools for biomedical text analysis, lexical resources, and manually annotated samples with semantic information has resulted in an explosion of event-related research in the domain (*cf.* <http://nactem.ac.uk/genia/>, <https://www.i2b2.org/>). Semantically driven literature analysis and literature-based knowledge discovery provide a lot of challenging research topics and a paradigm shift is taking place in the biomedical domain, from relation models in information extraction research to more expressive event models, *cf.* Björne *et al.* (2010).

The goal of this paper is to provide a snapshot of ongoing work that aims to develop and apply an appropriate infrastructure for automatic event extraction in the Swedish medical domain. In the long run, we are particularly interested in developing tools to support health care professionals and researchers to rapidly identify and semantically exploit relevant information in large textual repositories.

2. Theoretical Background

The FrameNet approach is based on the linguistic theory of frame semantics (Fillmore *et al.*, 2003) supported by corpus evidence. A semantic frame is a script-like structure of concepts which are linked to the meanings of linguistic units and associated with a specific event or state. Each frame identifies a set of frame elements, which are frame specific semantic roles (both core and non-core ones). Furthermore, roles may be expressed overtly, left unexpressed or not explicitly linked to the frame via linguistic conventions. In this work, we only deal with the first type of such roles. FrameNet documents the range of semantic and syntactic combinatory possibilities of frame evoking lexical units (LU), phrases and clauses by abstracting away from syntactic differences. A LU can evoke a frame, and its syntactic dependents can fill the frame element slots.

2.1 The Swedish FrameNet

The Swedish FrameNet (SweFN++) is a lexical resource under development, based on the English version of FrameNet constructed by the Berkeley research group. The SweFN++ is available as a free resource at <http://spraakbanken.gu.se/swefn/>. Compared to the Berkeley FrameNet, SweFN++ is expanded with information about the domain of the frames, at present the medical and the art domain. Since frame classification is based on general-domain frame semantics, several efforts have been described to domain adaptations even for English (Dolbey *et al.*, 2006). Medical frames in SweFN include: *Addiction*; *Cure*; *Recovery*; *Experience_bodily_harm*; *Falling_ill*; *Administration_of_medication* etc. For instance, the *Cure* frame describes a situation involving a number of core roles such as: *Affliction*, *Healer*, *Medication*, *Patient* etc., and is evoked by lexical units such as *detoxify*, *heal*, *surgery*, and *treat*.

3. Experimental Setting

The materials used so far for the extraction of relevant text samples and for aiding the recognition of relevant frame elements in the *Administration_of_medication* samples, using pattern matching approaches, include: *text samples*, taken from the MEDLEX corpus; (Kokkinakis, 2008); *medical terminology* (the Swedish nomenclature SNOMED CT); *list of medicines* (taken from FASS, the Swedish national formulary); *semi-automatic acquired drug/substance/disease lexicon extensions* (e.g. generic expressions of drugs and diseases, misspellings etc.); *lists of key words* (e.g. drug forms [pill, tablet, capsule], drug administration paths [intravenous, intravesical, subcutaneous], volume units [mg, mcg, IE, mmol] and various abbreviations and variants [iv, i.v., im, i.m. sc, s.c., po, p.o., vb, T]).

As a method we apply a simplistic rule-based approach (to be used as a baseline for future work) by performing three major steps: *pre-processing* (selecting 100 sample sentences of each frame using trigger words, i.e. relevant LUs and manual annotation of samples, [by the author]; Fig 1); *processing* (named entity, terminology and key word identification) and *post-processing* (e.g., modelling observed patterns in the rules, such as: `<Drug_name> <Drug_strength> <Frequency>`, normalization of labels and merging results). Through manual analysis of the annotated examples we get an approximation of how the examined medical events can be expressed in the data. This way we can model rules (regular expressions) for the task and also have annotated data for future planned supervised learning extensions. For the main processing step we apply named entity recognition which identifies and annotates relevant frame elements such as time expressions, various numerical information types and terminology. These annotations are important since they are both required by the frames and appear regularly in the context of the medical frames.

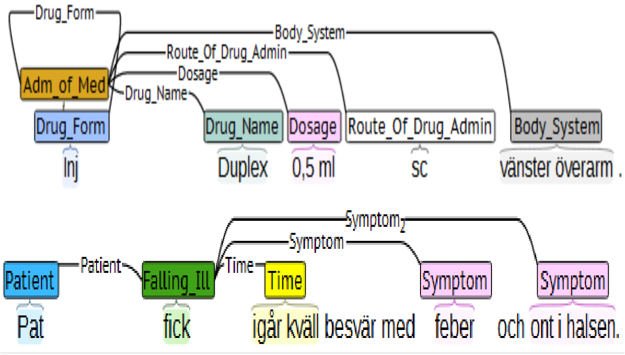


Figure 1. Examples of manually annotated data with the *Administration_of_Medication* (top) and the *Falling_Ill* (bottom) frames using *brat* (Stenetorp *et al*., 2012).

Table 1 shows the evaluation results (complete match) for the top-3 frame elements (most occurrences in a test set of 40 sentences). Some of the elements could not be found in the sample, while some had very few occurrences and this is the reason we chose not to formally evaluate all of them at this stage.

Frame Elements (#)	**Admin_of_Medic.**		**Falling_Ill**	
	Precision	Recall	Precision	Recall
Drug_Name (53)	93.8%	83.6%		
RouteOfDrugAdm (27)	100%	96.2%		
Dosage (21)	96.1%	94.2%		
Patient (39)			100%	92.8%
Ailment (33)			88.5%	91.1%
Symptom (20)			78.9%	82.1%

Table 1. Evaluation of the most frequent Frame Elements in a test set with their occurrences in parenthesis.

This vertical level evaluation assess the extraction of each frame element individually. Many problematic issues remain. For instance, certain elements are difficult to capture using regular expressions, i.e. `<Purpose>`, `<Outcome>` and `<Circumstance>.` These seem the most problematic since these element shows great variability and expressed by common language patterns. Perhaps syntactic parsing needs to be exploited.

4. Conclusions and Future Work

This paper has outlined current work towards natural language understanding and in particular event-based information extraction using frame semantics. We have been working with two of such frames and experimenting with simple pattern matching approaches in order to use as a baseline for future experiments. The driving force for the experiments is the theory of frame semantics, which allows us to work with a holistic and detailed semantic event description than it has been previously reported in similar tasks or in efforts using for instance most traditional methods based on relation extraction. Moreover, event extraction is more complicated and challenging than relation extraction since events usually have internal structure involving several entities as participants allowing a detailed representation of more complex statements. In the near future we intend to investigate the validity of the medical frames by manually annotating authentic samples for all available medical frames and also combine the pattern-based approaches with supervised learning for automatic extraction and labelling of frame elements.

5. Acknowledgements

This work was supported by the Centre for Language Technology at the University of Gothenburg and the Swedish Research Council project: *Swedish FramNet++ - a lexical infrastructure for Swedish language technology* (2011-13).

6. References

Ananiadou S., Pyysalo S., Tsujii J. and Kell DB. (2010). Event extraction for systems biology by text mining the literature. *Trends Biotechnol*. 28(7):381-90.

Björne J., Ginter F., Pyysalo S., Tsujii J. and Salakoski T. (2010). Complex event extraction at PubMed scale. *Bioinformatics* 15;26(12):i382-90.

Chambers N. and Jurafsky D. (2011). Template-based information extraction without the templates. Proc of the 49th Annual Meeting of ACL: HLT. Pp 976-986. Oregon, USA.

Dolbey A., Ellsworth M. and Scheffczyk J. (2006). BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. 2nd Workshop on Formal Biomed. Knowledge Repres. (KR-MED). Baltimore, USA.

Fillmore CJ., Johnson CR. and Petruck MRL. (2003). Background to FrameNet. *J of Lexicography*. 16(3).

Kokkinakis D. (2008). A Semantically Annotated Swedish Medical Corpus. 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.

Rau L., Krupka G., Jacobs P., Sider I. and Childs L. 1992. MUC-4 test results and analysis. 4th Message Understanding Conf. (MUC-4).

Ruppenhofer J. et al. (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. The NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Colorado, USA.

Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S. and Tsujii J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. Proc of the Eur. ACL. Pp. 102-107. Avignon, France.

Formal semantics for perception

Staffan Larsson

Göteborgs Universitet
Box 200, SE405 30 Göteborg
sl@ling.gu.se

Abstract

A representation linking subsymbolic perceptual aspects of meaning to logic-related aspects is proposed. We show how a simple classifier of spatial information based on the Perceptron can be cast in TTR (Type Theory with Records).

1. Introduction

In dynamic semantics, meanings are context-update functions which take an input context and return an updated (output) context. In this paper, a dynamic semantic approach to subsymbolic perceptual aspects of meaning is presented. We show how a simple classifier of spatial information based on the Perceptron can be cast in TTR (Type Theory with Records) (Cooper, 2012). A large variety of linguistic phenomena related to logical/symbolic meaning have already been addressed within this framework. Consequently, the TTR perceptron indicates that TTR may be a useful framework for integrating subsymbolic aspects of meaning in a way which allows us to keep around the accumulated insights from formal semantics.

2. The left-or-right game

As an illustration, we will be using a simple language game whose objective is to negotiate the meanings of the words “left” and “right”. A and B are facing a framed surface on a wall, and A has a bag of objects which can be attached to the framed surface. The following procedure is repeated:

1. A places an object in the frame
2. B orients to the new object, assigns it a unique individual marker and orients to it as the current object in shared focus of attention
3. A says either ”left” or ”right”
4. B interprets A’s utterance based on B’s take on the situation. Interpretation involves determining whether B’s understanding of A’s utterance is consistent with B’s take on the situation.
5. If an inconsistency results from interpretation, B assumes A is right, says “aha”, and learns from this exchange; otherwise, B says “okay”

3. Subsymbolic semantics

In this section, we will show how a TTR-based dynamic semantic account of meaning can be extended to incorporate subsymbolic aspects of meaning. Examples will be based on the left-or-right game introduced above.

3.1 Perceptual meanings as classifiers

We take the lexical meaning $[e]$ of an expression e to often contain not only compositional semantics but also perceptual meaning (at least for non-abstract expressions). By this we mean that aspect of the meaning of an expression which allows an agent to detect objects or situations referred to by the expression e. For example, knowing the perceptual meaning of “panda” allows an agent to correctly classify pandas in her environment as pandas. Likewise, an agent which is able to compute the perceptual meaning of “a boy hugs a dog” will be able to correctly classify situations where a boy hugs a dog. We can therefore think of perceptual meanings as classifiers of sensory input.

3.2 A TTR perceptron classifier

Classification of perceptual input can be regarded as a mapping of sensor readings to types. To represent perceptual classifiers, we will be using a simple perceptron. A perceptron is a very simple neuron-like object with several inputs and one output. Each input is multiplied by a weight and if the summed inputs exceed a threshold, the perceptron yields as output, otherwise 0 (or in some versions -1).

$$o(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} > \text{t} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{w}\cdot\mathbf{x} = \sum_{i=1}^{n} w_i x_i = w_1x_1 + w_2x_2 + \ldots + w_nx_n$

The basic perceptron returns a real-valued number (1.0 or 0.0) but when we use a perceptron as a classifier we want it to instead return a type. Typically, such types will be built from a predicate and some number of arguments; for the moment we can think of this type as a “proposition”.

A TTR classifier perceptron for a type P can be represented as a record:

$$\begin{bmatrix} \text{w} & = & \begin{bmatrix}0.800 & 0.010\end{bmatrix} \\ \text{t} & = & 0.090 \\ \text{fun} & = & \lambda v : \text{RealVector} \\ & & (\begin{cases} P & \text{if } v\cdot\text{w} > \text{t} \\ \neg P & \text{otherwise} \end{cases}) \end{bmatrix}$$

Where p.fun will evaluate to

$$\lambda v : \text{RealVector}$$
$$\left(\begin{cases} \text{P} & \text{if } v \cdot \begin{bmatrix}0.100 & 0.200\end{bmatrix} > 0.090 \\ \neg\ \text{P} & \text{otherwise} \end{cases}\right)$$

3.3 Situations and sensors

In the left-or-right game, we will assume that B's take on the situation includes readings from a position sensor (denoted "sr_{pos}") and a field foc-obj for an object in shared focus of attention. The position sensor returns a two-dimensional real-valued vector representing the horizontal vertical coordinates of the focused object: $\begin{bmatrix}x & y\end{bmatrix}$ where $-1.0 \leq x, y \leq 1.0$ and $\begin{bmatrix}0.0 & 0.0\end{bmatrix}$ represents the center of the frame.

Here is an example of B's take on the situation prior to playing a round of the left-or-right game:

$$\text{s}_1^B = \begin{bmatrix} \text{sr}_{pos}{=}\begin{bmatrix}0.900 & 0.100\end{bmatrix} & : & \text{RealVector} \\ \text{foc-obj=obj}_{45} & : & \text{Ind} \\ \text{spkr=A} & : & \text{Ind} \end{bmatrix}$$

In s_1^b, B's sensor is oriented towards obj_{45} and sr_{pos} returns a vector corresponding to the position of obj_{45}.

3.4 Utterance interpretation

We will take parts of the meaning of an uttered expression to be *foregrounded*, and other parts to be *backgrounded*. Background meaning (bg) represents constraints on the context, whereas foreground material (fg) is the information to be added to the context by the utterance in question. Both background and foreground meaning components are represented in TTR as types T_{bg} and T_{fg}.

The meaning of a sentence is modelled as a function from a record (representing the context) of the type T_{bg} specified by the background meaning, to a record type representing the type of the foreground meaning, T_{fg}.

$$\lambda r : T_{bg}(T_{fg})$$

When updating an agent's take on the context, given a current take on the context T, if $T \sqsubseteq T_{bg}$ (i.e., T is a subtype of T_{bg}, which informally means that T minimally contains the information specified by T_{bg} but possibly also other information) then the updated context T' is $T \underset{.}{\wedge} T_{fg}$ (but with any occurrences of bg in T_{fg} replaced by r). The $\underset{.}{\wedge}$ is a *merge* operator such that $T_1 \underset{.}{\wedge} T_2$ is T_1 extended with T_2.

$$\begin{bmatrix}\text{a=1:Int} \\ \text{b=2:Int}\end{bmatrix} \underset{.}{\wedge} \begin{bmatrix}\text{c=3:Int}\end{bmatrix} = \begin{bmatrix}\text{a=1:Int} \\ \text{b=2:Int} \\ \text{c=3:Int}\end{bmatrix}$$

3.5 The meaning of "right"

We can now say what a meaning in B's lexicon might look like. In our representations of meanings, we will combine the TTR representations of meanings with the TTR representation of classifier perceptrons. Agent B's initial take on the meaning of "right" is represented thus:

$$[\text{right}]^B =$$
$$\begin{bmatrix} \text{w} = \begin{bmatrix}0.800 & 0.010\end{bmatrix} \\ \text{t} = 0.090 \\ \text{bg} = \begin{bmatrix} \text{sr}_{pos} & : & \text{RealVector} \\ \text{foc-obj} & : & \text{Ind} \\ \text{spkr} & : & \text{Ind} \end{bmatrix} \\ \text{fg} = \begin{bmatrix} \text{c}^{perc}_{right} = \begin{bmatrix} \text{sr}_{pos} = \text{bg.sr}_{pos} \\ \text{foc-obj = bg.foc-obj} \end{bmatrix} : \\ \begin{cases} \text{right(bg.foc-obj)} & \text{if bg.sr}_{pos}\text{·w} > \text{bg.sr}_{pos}\text{·t} \\ \neg\text{right(bg.foc-obj)} & \text{otherwise} \end{cases} \end{bmatrix} \end{bmatrix}$$

The fields w and t specify weights and a threshold for a classifier perceptron which is used to classify sensor readings. The bg field represents constraints on the input context, which requires that there is a position sensor reading and a focused object foc-obj. In the fg field,the value of c^{perc}_{right} is a proof of either or right(foc-obj) or ¬right(foc-obj), depending on the output of the classifier perceptron which makes use of w and t. Here, right(y) is a perceptual "proposition" (a type constructed from a predicate), and objects of this type are proofs that y is (to the) right. As a proof of right(foc-obj) we count a "snapshot" of relevant parts of the situation, consisting of the current sensor reading and a specification of the currently focused object.

4. Contextual interpretation

Player A picks up an object and places it in the frame, and B finds the object and assigns it the individual marker obj_{45}, directs the position sensor to it and gets a reading. Player A now says "right", after which B's take on the situation is s_1^B (see above).

To interpret A's utterance, after checking that $\text{s}_1^B \sqsubseteq [\text{right}]^B\text{.bg}$, B computes $[\text{right}]^B\text{.fg} \underset{.}{\wedge}\ \text{s}_1^B$ to yield a new take on the situation s_2^B:

$$\text{s}_2^B = [\text{right}]^B \underset{.}{\wedge}\ \text{s}_1^B =$$
$$\begin{bmatrix} \text{sr}_{pos}{=}\begin{bmatrix}0.900 & 0.100\end{bmatrix}\text{:RealVector} \\ \text{foc-obj=obj}_{45}\text{:Ind} \\ \text{spkr=A:Ind} \\ \text{c}^{perc}_{right} = \begin{bmatrix} \text{sr}_{pos} = \begin{bmatrix}0.900 & 0.100\end{bmatrix} \\ \text{foc-obj = obj}_{45} \end{bmatrix} \text{:right(obj}_{45}\text{)} \end{bmatrix}$$

Here, the classifier takes s_1^B to contain a proof of right(obj_{45}). For an account of learning in the framework proposed above, see (Larsson, 2011).

5. References

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Staffan Larsson. 2011. The ttr perceptron: Dynamic perceptual meanings and semantic coordination. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, Los Angeles (USA).

A Bilingual Treebank for the FraCaS Test Suite

Peter Ljunglöf and Magdalena Siverbo

Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology
peter.ljunglof@gu.se

Abstract

We have created an open-source bilingual treebank for 99% of the sentences in the FraCaS test suite (Cooper et al., 1996). The treebank was built in conjunction with associated English and Swedish lexica written in the Grammatical Framework Resource Grammar (Ranta, 2009). The original FraCaS sentences are English, and we have tested the multilinguality of the Resource Grammar by analysing the grammaticality and naturalness of the Swedish translations. 86% of the sentences are grammatically and semantically correct and sound natural. About 10% can probably be fixed by adding new lexical items or grammatical rules, and only a small amount are considered to be difficult to cure.

1. The FraCaS corpus

The FraCaS textual inference problem set (Cooper et al., 1996) was built in the mid 1990's by the FraCaS project, a large collaboration aimed at developing resources and theories for computational semantics. The test set was later modified and converted to a corpus in XML format,[1] and it is this modified version that has been used in this project. The corpus consists of 346 problems each containing one or more statements and one yes/no-question. The total number of unique sentences in the corpus is 874.

The FraCaS problems are divided into 9 broad categories which cover many aspects of semantic inference. The categories are called *quantifiers*, *plurals*, *anaphora*, *ellipsis*, *adjectives*, *comparatives*, *temporal reference*, *verbs*, and *attitudes*, and they are also sub-categorised and sub-sub-categorised in an hierarchy of semantic phenomena. Each problem starts with one or more premises, and a question that can be answered with *yes*, *no* or *unknown*. Here is an example from the *ellipsis* category, with two different answers depending on whether the pronoun "*one*" refers to the "*red car*" or just the "*car*":

P: John owns a red car.
P: Bill owns a fast one.
Q: Does Bill own a fast red car?
A: Yes / Unknown.

2. Grammatical Framework

Grammatical Framework (GF) (Ranta, 2011) is a grammar formalism based on type theory. The main feature is the separation of abstract and concrete syntax. The abstract syntax of a grammar defines a set of abstract syntactic structures, called abstract terms or trees; and the concrete syntax defines a relation between abstract structures and concrete structures. The concrete syntax is expressive enough to describe language-specific linguistic features such as word order, gender and case inflection, and discontinuous phrases.

GF has a rich module system where the abstract syntax of one grammar can be used as a concrete syntax of another grammar. This makes it possible to implement grammar resources to be used in several different application domains. These points are exploited in the GF Resource Grammar Library (Ranta, 2009), which is a multilingual GF grammar with a common abstract syntax for 25 languages, including Finnish, Persian, Japanese and Urdu. The main purpose of the Grammar Library is as a resource for writing domain-specific grammars.

3. The English Grammar and Treebank

To be able to construct a GF treebank we need a grammar and a lexicon that can describe every sentence in the corpus. We have used the GF Resource Grammar as underlying grammar, and added lexical items that capture the FraCaS domain. On top of the resource grammar we have added a few new grammatical constructions, as well as functions for handling elliptic phrases.

In total, we used 107 grammatical functions out of the 189 that are defined in the resource grammar. In addition we added four new grammatical constructions that were lacking, and grammar rules for different elliptic phrases.

The lexicon has in total 531 entries, divided into 63 adjectives, 77 adverbials, 20 conjunctions/subjunctions, 34 determiners, 142 nouns, 19 numerals, 40 proper nouns, 15 prepositions, 12 pronouns, and 109 verbs.

3.1 Additions to the grammar

Four different grammatical constructions were added to the grammar. They consist of natural extensions to and slight modifications of existing grammar rules. An example of a grammar extension is the idiom "*so do I*" / "*so did she*".

The resource grammar cannot handle all kinds of conjunctions and elliptical phrases. In the FraCaS corpus there are 35 sentences with more advanced elliptical constructions. Examples include "*Bill did [...] too*", and "*Smith saw Jones sign the contract and [...] his secretary make a copy*". Our solution was to introduce elliptic phrases in the grammar, one for each grammatical category. E.g., the first example contains an elliptic verb phrase, and the second an elliptic ditransitive verb. To reduce ambiguity, each elliptic phrase is explicitly linearized into the string "*[...]*".

3.2 Coverage

Of the 874 unique sentences, 812 could be parsed directly with the Resource Grammar and the implemented lexicon,

[1] http://www-nlp.stanford.edu/∼wcmac/downloads/fracas.xml

	Total	% of sentences
Unique sentences	874	100%
Accepted by the RG	812	92.9%
- with grammar extensions	826	94.5%
- with elliptic phrases	860	98.4%
- with minor reformulation	866	99.1%
Unable to parse	8	0.9%

Table 1: Coverage of the English FraCaS grammar

No. parse trees	No. sentences	
1 – 9	598	69.1%
10 – 99	203	23.4%
100 – 999	49	5.7%
$\geq$ 1000	16	1.8%

Table 2: Ambiguity of the FraCaS treebank

as shown in table 1. With the three additional grammatical constructions 14 more sentences were parsed. The addition of elliptical phrases increased the number of sentences by another 34. Of the 14 remaining sentences, we could parse 6 more by doing some minor reformulations, such as moving a comma or adding a preposition.

All trees in the FraCaS treebank are implemented in the GF grammar described above. This grammar can be used by itself for parsing and analysing similar sentences. We parsed the 866 sentences covered by the grammar and counted the number of trees for each sentence. Table 2 shows that the grammar is moderately ambiguous, where almost 70% of the sentences have less than 10 different parse trees, and over 90% have less than 100 trees. The median is for a sentence to have 5 parse trees, and the largest number of trees for a sentence is 33,048.

Note that the number of parse trees are misleading for the 34 sentences with elliptic phrases, since ellipsis is linearised as "*[. . .]*" in the FraCaS grammar. If we had made the elliptic phrases invisible, the number of parse trees would increase dramatically.

4. The Swedish Corpus

As a first step towards making the treebank multilingual, we created Swedish translations of the sentences, by writing a new Swedish lexicon. Then we evaluated the translations and iteratively made changes to the trees to make the translations better. Note that since we use exactly the same syntax trees for the Swedish and English sentences, we had to make sure that the original English sentences were not changed when we modified the trees.

This means that we did not translate the English sentences manually, but instead we translated the lexicon and let the Swedish Resource Grammar take care of linearizing the treebank into Swedish. Currently, out of the 866 trees in the treebank, 748 are linearized into grammatically correct and comprehensible Swedish sentences.

4.1 Coverage

Table 3 gives an overview of the coverage of the Swedish lexicon and grammar. Of the 866 unique trees in the treebank, we consider 748 to have good Swedish translations. The remaining 118 sentences had some problems which we divided into five different classes – idioms, agreement, future tense, elliptical phrases, and more difficult errors. Of these 118 problematic Swedish sentences we believe that more than two thirds should be possible to add to the treebank without too much trouble.

	Total	% of sentences
Sentences in treebank	866	100%
Correct Swedish translation	748	86.4%
Problematic sentences	118	13.6%
– idioms	31	3.6%
– agreement	24	2.8%
– future tense	12	1.4%
– elliptical	19	2.2%
– uncomprehensible	32	3.7%

Table 3: Coverage of the Swedish FraCaS grammar

5. Conclusion

The FraCaS treebank was created in 2011 as a small project financed by the Centre for Language Technology (CLT) at the University of Gothenburg. The project used less than three person months to create a treebank for the FraCaS test suite, together with a bilingual GF grammar for the trees. The coverage of the English grammar is 95–99%, depending on whether you include elliptic phrases or not. The Swedish grammar has a coverage of 86%.

The making of this treebank has been a stress test, both for GF and for the resource grammar. The main work in this project has been performed by a person who is an experienced computational linguist, but had never used GF before. This means that the project has been a test of how easy it is to learn and start using GF and its resource grammar. Furthermore, it was a test of the coverage of the existing grammatical constructions in the resource grammar.

The treebank is released under an open-source license, and can be downloaded as a part of the Gothenburg CLT Toolkit.[2] There is also a technical report describing the treebank in more detail (Ljunglöf and Siverbo, 2011).

6. References

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jaspars Jan, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Deliverable D16, FraCaS Project.

Peter Ljunglöf and Magdalena Siverbo. 2011. A bilingual treebank for the FraCaS test suite. CLT project report, University of Gothenburg.

Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

[2] Available from URL http://www.clt.gu.se/clt-toolkit

A Chunking Parser for Semantic Interpretation of Spoken Route Directions in Human-Robot Dialogue

Raveesh Meena, Gabriel Skantze, Joakim Gustafson
Department of Speech, Music and Hearing
KTH, Stockholm
raveesh@csc.kth.se, gabriel@speech.kth.se, jocke@speech.kth.se

Abstract

We present a novel application of the *chunking parser* for data-driven semantic interpretation of spoken route directions into *route graphs* that are useful for robot navigation. Various sets of features and machine learning algorithms were explored. The results indicate that our approach is robust to speech recognition errors, and could be easily used in other languages using simple features.

1. Introduction

It is desirable to endow urban robots with spoken dialogue capabilities so that they can seek route directions from passersby to navigate their way in unknown surroundings. To understand freely spoken route directions a robot's dialogue system would require a spoken language understanding (SLU) component that (i) is *robust* in handling automatic speech recognition (ASR) errors, (ii) learns *generalization* to deal with unseen concepts in free speech, and (iii) preserve the highly *structured relations* among various concepts in a route direction instruction. Existing approaches to SLU in dialogue systems do not cater to one or the other of these requirements.

Grammar based parsers for semantic interpretations of text, besides requiring envisaging all the combinatory rules, are not robust in handling ASR errors. While *keyword spotting* based approaches to SLU has been useful in form-filling task domains, the semantics of route directions is highly structured and cannot be treated as simple "bag of words or concepts". For example, route instruction "*after the church turn left*" contains not just the concepts AFTER, CHURCH, TURN and LEFT but also the structural relations among them: that the action of turning to left has to be taken only after the church.

Earlier in Johansson et al. (2011), we presented a data-driven approach to semantic interpretation of manual transcriptions of route instructions, given in Swedish, into *conceptual route graphs* (CRGs) that represent the semantics of human route descriptions. In Meena et al. (2012) we applied this approach for semantic interpretation of spoken route directions given in English. The results indicate that our approach is robust in handling ASR errors. Here we present an overview of the work reported in Meena et al. (2012).

2. Chunking parser for semantic interpretation

Our approach (Johansson et al., 2011) to automatically interpret manual transcriptions of route instructions into CRGs (Müller et al., 2000), is a novel application of Abney's *chunking parser* (Abney, 1991). In a CRG, e.g. Figure 1, the nodes represent the semantic concepts and the edges their attributes. The concepts, their attributes and argument types are defined in the type hierarchy of the domain model using the specification in the JINDIGO dialogue framework (Skantze, 2010).

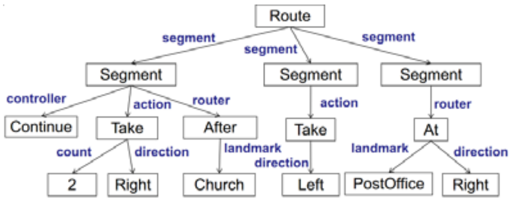


Figure 1: The conceptual route graph for the route instruction "*go straight and take the second right after the church then eh take a left the post is on the right hand side.*"

We apply the *Chunker* stage of the chunking parser for finding base concepts in a given sequence of words. For example, route instruction "*turn left after eh the church*" could be chunked as the following:

[**ACTION** *turn*] [**DIRECTION** *left*] [**ROUTER** *after*] [**FP** *eh*] [**LANDMARK** *the church*]

To turn chunking into a classification problem, we followed the common practice of assigning two labels for each type of chunk: one with prefix B- for the first word in the chunk and one with prefix I- for the remaining words. The *Attacher* then takes a base concept (a chunk) as input and does two things: First, it may assign a more specific concept class (like CHURCH). To allow it to generalize, the Attacher also assigns all ancestor classes, based on the domain model (i.e. BUILDING for CHURCH; this, however, is not shown in the example). The second task for the Attacher is to assign attributes, e.g. *direction*, and assign them values, e.g. →, which means that the interpreter should look for a matching argument in the right context. While the Chunker in our approach is a *single-label classifier* the Attacher is a *multi-label classifier* where none, one or several labels may be assigned to a chunk. The Chunker output from above could be modified by the Attacher as the following:

[**TAKE** (*direction*: →) *turn*] [**RIGHT** *right*] [**AFTER** (*landmark*: →) *after*] [**DM** *eh*] [**CHURCH** *the church*]

As a final step, heuristic rules were used to group the CONTOLLER, ROUTER and ACTION chunks into *route segments*, which collectively form a CRG. To measure the performance of the Chunking parser we used the notion of Concept Error Rate (CER), which in our approach is the weighted sum of the edits required in the reference CRG key to obtain the resulting CRG.

3. Chunking parser for SLU

In Meena et al. (2012) we made three extensions to this approach. First, we introduced another chunk learner – the *Segmenter* – to automatically learn *route segments* in a sequence of chunks. The Chunker output shown earlier could be segmented as the following:

[SEGMENT [ACTION *turn*] [DIRECTION *right*] [ROUTER *after*] [FP *eh*] [LANDMARK *the church*] **]**

The Attacher performs the same tasks as earlier, except that it now looks for attachments only within the route segment. Secondly, we verified whether our approach could be applied for semantic interpretation of route directions given in another language. Third, we evaluated the performance of the Chunking parser on ASR results. Towards this, we used the IBL corpora of route instructions (Kyriacou et al., 2005). It contains audio recordings and manual transcriptions of 144 spoken route instructions given in English. As a first step, we evaluated the Chunking parser's performance on manual transcriptions and obtained the baseline for comparing its relative performance on ASR results. 30 route instructions were manually annotated and used as the cross-validation set. Next, we trained an off-the-shelf ASR system with the remaining 113 route instructions. The best recognized hypothesis for each instruction (in the cross-validation set) obtained through this trained ASR was used for validating the Chunking parser's performance.

We tested the Chunking parser's performance on Naïve Bayes (NB) and Linear Threshold Unit (LTU) algorithms. Two types of LTU: Sparse Perceptron (SP) and Sparse Averaged Perceptron (SAP) were tested. Due to space constraints we present, in the first column of Table 1, only those feature (combinations) for which the learners obtained the best results. Interested readers are referred to Meena et al. (2012) for complete details.

4. Results

The performance scores of a *keyword spotting* based method for *Chunking parser* were used as the baseline for drawing comparisons. A baseline CER of 50.83 was obtained for the Chunker using the NB learner. For the Segmenter a CER of 77.22 was obtained using the NB learner, and for the Attacher a CER of 75.07 was achieved with the SAP learner.

In general LTUs performed better than NB algorithms. The best performance for the Chunker is obtained using the SAP learner in conjunction with the additive features shown in Table 1-A. The Segmenter performed best using the SAP (cf. Table 1-B). The Attacher performed best using the SAP learner (cf. column 2, Table 1-C); however, due to poor placement of route segment boundaries by the Segmenter, it could not always attach concepts to their valid argument(s). To obtain an estimate of Attacher's performance independent of the Segmenter we compared only the sub-graphs in all the *route segments* of a CRG with their counterpart in the reference CRG (cf. column 3, Table 1-C).

A: Chunker performances with additive features.			
Features	**CER_{NB}**	**CER_{SP}**	**CER_{SAP}**
Word instance	50.83	46.15	45.17
+Word window	17.31	21.33	20.82
+Previous tags	18.16	10.86	**10.64**
B: Segmenter performances with the best training feature.			
Features	**CER_{NB}**	**CER_{SP}**	**CER_{SAP}**
Chunk label window	31.67	**25.83**	28.89
C: Attacher performances with the best training feature.			
Features	**CER_{SP}**	**CER_{SAP}**	**rgCER_{SAP}**
Bag of words (BW)	29.42	**29.11**	**19.99**

Table 1: *Performances of the Chunking parse components.*

Figure 2 illustrates the Chunking parser performances (CER) w.r.t to ASR performances in recognizing the spoken route instructions. The dashed-red line is Chunker parser's baseline performance on the manual transcriptions. The CER curves suggest that Chunking parser's performance follows WER.

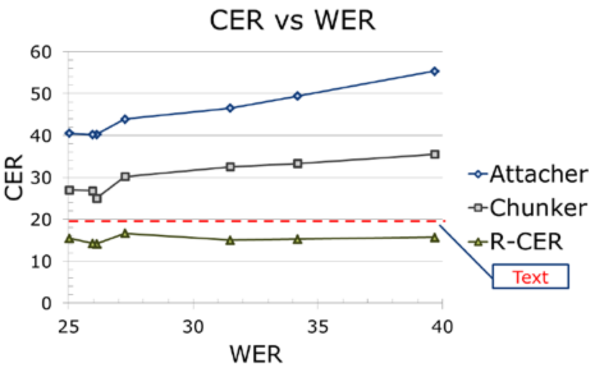


Figure 2*: Chunking parser's CER w.r.t ASR's WER (R-CER is relative-CER = CER minus WER)*

The rather steady *relative* CER: R-CER (the relative gain in CER due to ASR errors) in Figure 2 highlights the robustness of our approach in dealing with errors in speech recognition. In addition to this, Chunking parser's performance on transcribed route instructions given in Swedish (CER 25.60 (Johansson et al., 2011)) and in English (CER of 19.99 vs. baseline CER of 77.70) are encouraging figures that indicate that our approach can be easily used in other languages using simple features.

5. References

Abney, S. (1991). Parsing by chunks. In Berwick, R. C., Abney, S. P., & Tenny, C. (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics* (pp. 257-278). Dordrecht: Kluwer.

Johansson, M., Skantze, G., & Gustafson, J. (2011). Understanding route directions in human-robot dialogue. In *Proceedings of SemDial* (pp. 19-27). Los Angeles, CA.

Kyriacou, T., Bugmann, G., & Lauria, S. (2005). Vision-based urban navigation procedures for verbally instructed robots. *Robotics and Autonomous Systems*, *51*(1), 69-80.

Meena, R., Skantze, G., & Gustafson, J. (2012). A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue. In *Proceedings of Interspeech*. Portland, OR, US.

Müller, R., Röfer, T., Lankenau, A., Musto, A., Stein, K., & Eisenkolb, A. (2000). Coarse qualitative descriptions in robot navigation. In Freksa, C., Brauer, W., Habel, C., & Wender, K-F. (Eds.), *Spatial Cognition II* (pp. 265-276). Springer.

Skantze, G. (2010). *Jindigo: a Java-based Framework for Incremental Dialogue Systems*. Technical Report, KTH, Stockholm, Sweden.

Improving the acquisition of synonymy relations with contextual filtering: a preliminary study

Mounira Manser, Thierry Hamon

LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny Cedex, France
manser.mounira@gmail.com, thierry.hamon@lipn.univ-paris13.fr

Abstract

The automatic acquisition of semantic relations between terms in specialised texts is generally based on the exploitation of their internal structure or of their context. But it has been observed that, given the strategy, the proposed approaches suffer from a lack of quality either in precision or in recall. We tackle this problem by taking into account both the context and the internal structure of the synonymous terms. Our study shows that the proposed method is efficient: most of the synonymy relations validated by an expert are useful to infer synonymy relations while those rejected by the expert are not useful.

1. Introduction

Information retrieval and extraction in specialised corpora require domain specific structured information. The terminological resources partially answer to this need by providing the terms and semantic relations. But the acquisition of additional semantic relations between terms is required for a better coverage of the used corpora (Cabré, 1999; Spasic et al., 2005). NLP approaches dedicated to this objective usually exploit either the internal structure of terms (Jacquemin, 1999; Grabar and Zweigenbaum, 2000; Hamon and Nazarenko, 2001) or their contexts (Hearst, 1992; Curran, 2004). These works show a great variety in the proposed methods. Comparing their results, we have observed that the former approaches suffer from a low precision while the latter have a low recall. However, to our knowledge up to now no work proposed to combine internal analysis of terms with exploitation of their context to acquire semantic relations. To tackle this problem, we propose to take into account the context of synonymous terms together with their internal structure. More precisely, we propose to take the synonymy relations acquired with internal approach and to automatically filter them by taking into account the use of the terms in corpora.

2. Corpus-based semantic relation filtering

Our study is based on an internal approach for the acquisition of synonymy relations between multi-word terms (Hamon and Nazarenko, 2001), which assumes that initial synonymy relation between words can be propagated to multi-word terms through the compositional principle (for instance, given the synonymy relation between the words *pain* and *ache*, the terms *muscle pain* and *muscle ache* are inferred as synonyms). The inverse approach has been defined to induce initial relations from the relations issued from a terminology (Hamon and Grabar, 2008). (Hamon and Nazarenko, 2001) has shown that the quality of the inferred relations depends on the origin of the initial relations: the use of domain-specific relations give better precision but at the expense of the recall. However, given particular utterances, which can be captured through the context or the terminological use in corpora, semantic shifts of the relationship can be observed during the inference or the induction. Thus, even if the method has a good quality, the induced relations have to be filtered or contextualised (Hamon et al., 2012). Moreover, as the synonymy is a contextual relationship (Cruse, 1986), it is important to consider the context during the acquisition of the initial relations.

In that respect, we automatically filter the induced initial relations according to their usefulness in a corpus: the occurrence of the related words in a text and the use of the relations in the synonymy inference are the clues of their relevancy. The initial relations induced from a terminology are then exploited on corpora to infer relations between multi-word terms. We study initial and inferred relations, expecting that the initial relations exploited during inference are relevant while the initial relations rejected by an expert are not exploited in corpora.

3. Results and Discussion

We perform experiments with the 3,707 initial relations induced from the 101,254 synonymy relations proposed by Gene Ontology[1]. An expert validated 72% of them as synonymy. These relations have been filtered thanks to the inference of semantic relations between multi-word terms on five corpora of Medline[2] abstracts: Genia corpus[3] (2,000 abstracts, 400,000 words), BioNLP2011 corpus[4] (800 abstracts issued from the Genia corpus, 176,146 words), APOE corpus (1,580 abstracts associated to the apolipoprotein E gene, 346,339 words), Cael corpus (347 abstracts associated to the species Caenorhabditis elegans, 71,000 words), and Hosa corpus (355 abstracts associated to the species Homo sapiens, 75,379 words). The terms have been extracted with YaTeA[5].

Table 1 presents the results on the relation inference in the perspective of filtering of the initial relations: we consider the total number of all the inferred relations but also those issued only from initial relations validated as synonymy. We indicate the number of the initial relations used for inference (all the relations and only those validated as

[1] http://www.geneontology.org/

[2] http://www.ncbi.nlm.nih.gov/Entrez/

[3] http://www.nactem.ac.uk/genia/genia-corpus

[4] http://2011.bionlp-st.org/

[5] http://search.cpan.org/ thhamon/Lingua-YaTeA/

Corpora	Initial relations...					Inferred Relations...		
	Total		... validated as synonymy			Total	... from initial relations,	Precision
	Rel.	Productivity	Rel.	Precision	Productivity	rel.	validated as synonymy	
Genia	187	4.28	57	30.5	6.67	800	380	47.5
BioNLP	95	2.92	35	36.8	3.74	277	131	47.3
APOE	30	1.80	14	46.7	2.29	54	32	59.3
Cael	26	1.42	9	34.6	1.33	37	12	32.4
Hosa	21	1.24	9	42.9	1.22	26	11	42.3

Table 1: Results of the relation inference given initial relations, on the five corpora (the left part of the table describes the inference results, while the right part provides quantitative information about the initial relations; productivity of the initial relations provides information about the number of inferred relations from each initial relation)

synonymy), the productivity of those initial relations (the number of inferred relations for each initial relation), and proportion of synonymy relations among the initial and inferred relations.

The usefulness of the induced relations on the corpora is subject to a large variation. On the Genia corpus, 800 relations were inferred from 187 initial relations with a productivity of 4.28 relations, while when only the synonymy relations are considered, the productivity is 6.67. On BioNLP corpus, the productivity is much lower (2.92 and 3.74) while the proportion of synonymy relations is quite similar. A first analysis of these inferred relations shows that they are semantically correct. On the other corpora, even if the percentage of synonymy relations is similar or superior, few relations are inferred. Compared to the results from the Genia corpus, the number of relations is very low proportionally (Cael and Hosa corpora), even when the corpus size is similar (APOE corpus). The content of the corpus seems to have an influence on the number of the inferred relations.

These experiments show that initial relations validated as synonymy are more productive and proportionally more useful to infer relations when the corpus is large enough and when the resource content is adapted to the corpus. The low number of inferred relations compared to the number of initial relations can be explained by the fact that there are few semantic variations within the abstract corpora. Our hypothesis of exploiting the use of the terms in corpus to filter the initial relations seems to be correct. However, filtering by the use in corpus is not sufficient: it is a clue among others. It can be exploited to reinforce some initial relations or to give a more important weight to the relations used in corpus.

4. Conclusion and Perspectives

Considering the need to improve the quality of existing approach for induction of initial synonymy relations (Hamon and Grabar, 2008), we propose a method for filtering these relations according to their usefulness for inference of relations between multi-word terms. The results show that the initial relations validated as synonymy are mostly used to acquire the relations between complex terms. The use of terms in corpora seems to be an important clue for improving the previously proposed methodology for the induction of initial semantic relations. The study of the results also showed that the size and content of the corpus are to be taken into account. This observation will be confirmed by achieving the same type of experiments on corpora of full scientific articles collected on PubMed Central[6]. The main perspective to our work is the exploitation of the semantic information in the context of the terms. It will lead to a more complete description of the contexts in which the acquired synonymy relations are relevant. For instance, we aim at constraining the scope of application of initial relations using semantic categories associated to words and terms and machine learning approaches.

5. References

Maria Teresa Cabré. 1999. *Terminology. Theory, methods and applications*, volume 1 of *Terminology and Lexicography, Research and practice*. John Benjamins.

David A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

James R. Curran. 2004. *From distributional to semantic similarity*. Phd thesis, University of Edinburgh.

N. Grabar and P. Zweigenbaum. 2000. Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO-2000*, pages 765–784.

Thierry Hamon and Natalia Grabar. 2008. Acquisition of elementary synonym relations from biological structured terminology. In *Proceedings of CICLing*, pages 40–51.

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: experiment and results. In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

Thierry Hamon, Christopher Engström, Mounira Manser, Zina Badji, Natalia Grabar, and Sergei Silvestrov. 2012. Combining compositionality and pagerank for the identification of semantic relations between biomedical words. In *Proceedings of BioNLP Workshop*, pages 109–117.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the COLING'92*, pages 539–545.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL'99*, pages 341–348.

Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, September.

[6] http://www.ncbi.nlm.nih.gov/pmc/

SUC-CORE: SUC 2.0 Annotated with NP Coreference

Kristina Nilsson Björkenstam and Emil Byström

Computational Linguistics, Department of Linguistics
Stockholm University
kristina.nilsson@ling.su.se

Abstract

SUC-CORE is a subset of Stockholm Umeå Corpus 2.0 and Swedish Treebank, annotated with noun phrase coreference. While most coreference annotated corpora consist of texts of similar types within related domains, SUC-CORE consists of both informative and imaginative prose and covers a wide range of literary genres and domains.

1. Introduction

SUC-CORE is a 20 000 word subset of the Stockholm-Umeå Corpus (SUC 2.0) annotated with coreference relations between noun phrases. This subset consists of the same documents as the evaluation set of the Swedish Treebank.[1] Thus, the coreference annotation of SUC-CORE can be combined with the part-of-speech tagging, morphosyntactic analysis and named entity annotation of SUC 2.0 (Gustafson-Čapková and Hartmann, Eds., 2006), and the syntactic analysis of the Swedish Treebank (Nivre et al., 2008).

Through SUC-CORE, we offer annotated data for development and evaluation of coreference resolution for Swedish. To our knowledge, this is the only Swedish corpus with coreference annotation available for research.

2. Data

SUC 2.0 is a balanced corpus, covering various text types and stylistic levels. It is modeled on the Brown Corpus (Bonelli and Sinclair, 2006) and similar sample corpora with two main categories of texts, *informative prose* and *imaginative prose*. The first category consists of e.g., news text, feature articles, and scientific papers, and the second category of different genres of fiction. SUC 2.0 follows the general layout of Brown with 500 samples of text with a length of about 2,000 words each. These text samples are composed of excerpts from longer texts or a selection of short texts (Gustafson-Čapková and Hartmann, Eds., 2006).

SUC-CORE includes both informative and imaginative text of different genres and domains. The informative prose category consists of six files with foreign and domestic news texts and editorials from national and regional morning dailys, magazine articles on interior design, a textbook excerpt on biology, and an academic essay. The imaginative prose section includes excerpts from four novels of different genres (see table 1). Thus, SUC-CORE can be used for development and evaluation against different types of text.

Most comparable data sets annotated with coreference consist of texts of similar types within the same (or related) domains, e.g., newswire text in MUC-6 and MUC-7 (Hirschman and Chinchor, 1997), and newswire, broadcast transcripts, and blogs in ACE (Doddington et al., 2004).[2]

3. Coreference Annotation

(Van Deemter and Kibble, 1999) define coreference as a relation between two referring expressions α_1 and α_2. Assuming that both expressions are NPs and that both have a unique reference in the context in which they occur, α_1 and α_2 corefer if Reference(α_1) = Reference(α_2). During annotation, our goal has been to ensure that the resulting equivalence classes (or coreference chains) consist of NPs with identical references.

3.1 Entities and mentions

Following ACE (Doddington et al., 2004), we define an *entity* as an object or a set of objects in the world, and a *mention* as an expression which refers to an entity. The annotation task is restricted to three types of referring expressions:

- Name mentions (NAM): proper names and other named entities, e.g., *Robert Mugabe*. Tight appositions are included, e.g., *president* in the mention *president Mugabe*.
- Nominal mentions (NOM): NPs with a lexical noun, e.g., *partiets* (‘the party’), or a nominalized adjective or a participle as head, e.g., *den gamle* (‘the old+masc’).
- Pronominal mentions (PRO) consist of personal pronouns (e.g., *jag* (‘I’), *hon* (‘she’)), demonstrative pronouns (e.g., *denna* (‘this+uter (one)’), *detta* (‘this+neuter (one)’)), and reflexive pronouns (e.g., *sig* (‘himself’/‘herself’/‘itself’)). We also include possessives and genitives in this category.

Coordinated NPs are marked as one mention when this mention is coreferent with e.g., a plural pronoun. Plural anaphors with split antecedents are not annotated because the antecedents cannot be marked as one mention and thus yield an equivalence class where the NPs have non-identical references.

3.2 Annotation process

Using BRAT, a web-based annotation tool (Stenetorp et al., 2012), the data was annotated by the authors in collaboration, followed by a discussion of disagreements and final editing by the first author.

[1] STB. http://stp.ling.uu.se/∼nivre/swedish_treebank/

[2] Available through LDC. http://www.ldc.upenn.edu/

File	Genre	Source	Tokens	Mentions	Relations
I: Informative prose					
aa05	Press; political reportage* (foreign)	National daily	2056	713	235
aa09	Press; political reportage* (foreign, domestic)	Regional daily	2073	658	167
ba07	Press; editorials*	Regional	2100	670	118
ea10	Skills, trades and hobbies (interior design)*	Periodical	2194	734	154
ea12	Skills, trades and hobbies (biology)	Textbook	2017	715	178
ja06	Learned and scientific writing (humanities)	Textbook	2123	755	132
Total I:			**12563**	**4245**	**984**
II: Imaginative prose					
kk14	Fiction (Tunström, G. “Det sanna livet”)	Novel	2067	743	343
kk44	Fiction (Thorvall, K. “När man skjuter arbetare”)	Novel	2016	632	253
kl07	Crime (Nesser, H. “Det grovmaskiga nätet”)	Novel	2008	664	327
kn08	Romance (Dagsås, J. “Riddaren i mina drömmar”)	Novel	2004	617	330
Total II:			**8095**	**2656**	**1253**
TOTAL I + II:			**20658**	**6901**	**2237**

Table 1: Overview of SUC-CORE: file, genre, source, no. of tokens, mentions, and coreference between mentions. Files marked with (*) consist of selections of texts.

3.3 Annotation format

Each mention is annotated with a mention type (NAM, NOM, PRO) and connected to a specific span of text through character offsets. In the example below (from file aa05), the following mentions are identified: *Robert Mugabe*, *hans Zanuparti* (lit. ‘his Zanu party’), *hans* (‘his’), *partiets stängda dörrar* (lit. ‘the party’s closed doors’), *partiets* (‘the party’).

> ... officiellt jublade **Robert Mugabe** och **hans Zanuparti** på söndagen. Men bakom **partiets stängda dörrar** måste ...

During annotation, the mention *Robert Mugabe* with index T5 is marked as mention type NAM and connected to the span 154 to 167 in the source text.

```
...
T5   NAM   154   167   Robert Mugabe
T6   NAM   172   186   hans Zanuparti
T7   NOM   212   235   partiets stängda dörrar
T8   PRO   172   176   hans
T9   NOM   212   220   partiets
...
```

The coreference annotation is listed as pairwise relations between mentions, e.g., T5 (*Robert Mugabe*) and T8 (*hans*, ‘his’). This directed relation can link a mention to a previous or to a subsequent mention.

```
R1   Coref   Anaphora:T8   Antecedent:T5
R2   Coref   Anaphora:T9   Antecedent:T6
...
```

This format is similar to the BioNLP Shared Task standoff format.[3] We refer the reader to the documentation of SUC-CORE and the BRAT website[4] for further details.

4. Distribution

SUC-CORE is distributed by the Linguistics Department at Stockholm University.[5] SUC 2.0 and Swedish Treebank are distributed by Språkbanken at Gothenburg University.[6]

5. Concluding remarks

SUC-CORE covers a wide range of genres and domains. This allows for exploration of coreference across different text types, but it also means that there are limited amounts of data within each type. Thus, future work on coreference resolution for Swedish should include making more annotated data available for the research community.

6. References

E.T. Bonelli and J. Sinclair. 2006. Corpora. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 206–220. Elsevier, Oxford, second edition.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. In *Proceedings of LREC’04*.

S. Gustafson-Čapková and B. Hartmann, Eds. 2006. SUC 2.0. Department of Linguistics, Stockholm University.

L. Hirschman and N. Chinchor. 1997. MUC-7 Coreference Task Definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

J. Nivre, B. Megyesi, S. Gustafson-Čapková, F. Salomonsson, and B. Dahlqvist. 2008. Cultivating a swedish treebank. In *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, pages 111–120. Acta Universitatis Upsaliensis.

P. Stenetorp, S. Pyysalo, G. Topi, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Sessions at EACL 2012*, France, April. ACL.

K. Van Deemter and R. Kibble. 1999. What is coreference, and what should coreference annotation be? In A. Bagga, B. Baldwin, and S. Shelton, editors, *Proceedings of the ACL Workshop on Coreference and Its Applications*, Maryland, June. ACL.

[3] BioNLP. http://conll.cemantix.org/2011/data.html

[4] BRAT. http://brat.nlplab.org

[5] DALI, SU. http://www.ling.su.se/english/nlp/resources

[6] Språkbanken. http://spraakbanken.gu.se/

Properties of phoneme N-grams across the world's language families

Taraka Rama, Lars Borin

Språkbanken, Department of Swedish,
University of Gothenburg, Sweden
taraka.rama.kasicheyanula@gu.se, lars.borin@gu.se

Abstract

In this article, we investigate the properties of phoneme N-grams across half of the world's languages. The sizes of three different N-gram distributions of the world's language families obey a power law. Further, the N-gram distributions of language families parallel the sizes of the families, which also follow a power law distribution. The correlation between N-gram distributions and language family sizes improves with increasing values of N. The study also raises some new questions about the use of N-gram distributions in linguistic research, which we hope to be able to investigate in the future.

1. Introduction

There are about 7000 languages (Lewis, 2009) in the world which are classified into more than 120 language families. A language family is a group of related languages (or a single language when there are no known related languages) which have been shown to have stemmed from a common ancestor (Campbell and Poser, 2008). Each of these language families is assigned a tree structure in at least two classifications (Lewis, 2009; Hammarström, 2010). It has been shown that the sizes of the world's language families follow a power law distribution (Wichmann, 2005), in linguistics often referred to as a Zipfian distribution. The size of a language family is defined as the number of related languages included in the family. In this paper, we find that the rank plot of the size of phoneme N-grams for 45 language families follows a power law distribution as given in Figure 1. This finding is in parallel to that of Wichmann (2005).

2. Database

A consortium of international scholars known as ASJP (Automated Similarity Judgment Program; Wichmann et al. 2011) have collected reduced word lists (40 items from the original 200 item Swadesh word lists, selected for maximal diachronic stability) for more than half of the world's languages and embarked on an ambitious program for investigating automated classification of the world's languages. The ASJP database in many cases includes more than one word list for different dialects of a language (identified through its ISO 693-3 code). A word list is included into the database if it has attestations of at least 28 of the 40 items (70%).

Only language families with at least 4 members are included in our experiments. This leaves a dataset with 45 language families representing 3151 languages (or 4524 word lists) of the world. The names of the language families are as defined in the *Ethnologue* (Lewis, 2009). A word list might include known borrowings marked as such and they are subsequently removed from our experiments. The words in the ASJP database are transcribed using a reduced phonetic transcription known as *ASJP code* consisting of 34 consonants, 7 vowels and three modifiers which are used to combine the preceding segments. All click sounds are reduced to a single click symbol and tone distinctions are ignored. The phoneme N-gram profile for a language family is extracted as follows.

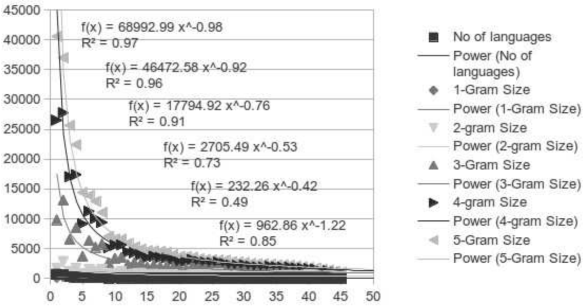


Figure 1: Power law distribution of phoneme N-grams

3. Method and results

All the word lists belonging to a single language family are merged together. Recall that the database can include more than one word list for a language as identified by a unique ISO 639-3 code. All the consecutive symbol sequences of length varying from 1–5 are extracted and the size of the N-gram profile is defined as the total number of unique 1–N-grams obtained through this procedure. Thus, a 3-gram profile consists of all the phoneme 1-, 2- and 3-grams. The size of the 3-, 4- and 5-gram profiles for each of the language families as defined in the Ethnologue is given in Table 1. Only the 3-, 4- and 5-gram profiles are provided, since those are the N-gram distributions with a higher goodness-of-fit. As evident from Figure 1, each of the N-gram profiles follows a power law. When a power law regression is applied to each of the distributions, the goodness-of-fit R^2 is 0.91, 0.96 and 0.97 for 3-grams, 4-grams and 5-grams respectively. The R^2 value of both 1-grams and 2-grams is quite low, only 0.49 and 0.73 when compared to the R^2 value of the number of languages, 0.85. The R^2 scores in Figure 1 for 4–5 grams are very high and fall within the range of the

Table 1: The number of languages (NOL), 3-gram, 4-gram and 5-gram profiles for 45 language families

Language family	NOL	3-gram	4-gram	5-gram	Language family	NOL	3-gram	4-gram	5-gram
Austronesian	692	9808	26527	40561	Macro-Ge	20	1813	2684	3180
Niger-Congo	615	13085	27766	36987	Sepik	22	1677	2623	3144
Trans-New Guinea	275	6492	17051	25634	Tai-Kadai	42	2036	2723	3077
Afro-Asiatic	201	8456	17403	22403	Chibchan	16	1522	2484	3035
Australian	104	3690	9143	14401	West Papuan	14	1244	2203	2806
Indo-European	139	6320	11252	13896	Eastern Trans-Fly	4	1300	2166	2729
Nilo-Saharan	123	5224	10046	12891	Dravidian	24	1376	2238	2674
Sino-Tibetan	137	5753	9386	11043	Lakes Plain	17	1142	2010	2518
Arawakan	39	2626	5148	7021	Border	8	1132	1922	2467
Austro-Asiatic	82	3370	5552	6608	South-Central Papuan	6	1074	1878	2464
Oto-Manguean	69	3522	5607	6579	Penutian	14	1326	2017	2384
Uto-Aztecan	43	2318	4395	5873	Panoan	15	1192	1915	2288
Altaic	46	2634	4304	5248	Witotoan	7	1185	1847	2264
Salishan	16	2492	3944	4903	Hokan	14	1253	1864	2192
Algic	19	1922	3466	4643	Quechuan	22	1101	1734	2093
Tupi	46	2250	3722	4619	Siouan	11	1131	1674	1952
Torricelli	21	2011	3518	4523	Na-Dene	15	1225	1637	1810
Mayan	48	2083	3485	4386	Hmong-Mien	15	1246	1563	1717
Tucanoan	18	1880	3162	3979	Totonacan	10	679	1139	1510
Ramu-Lower Sepik	14	1491	2738	3676	Khoisan	12	995	1265	1377
Carib	20	1662	2868	3649	Sko	12	775	1068	1179
North Caucasian	29	2180	3158	3537	Mixe-Zoque	12	625	897	1028
Uralic	23	1896	2818	3284					

correlation of 0.957 (with language family size), reported by Wichmann (2005).

4. Discussion and conclusions

As we have shown above, the correlation of N-gram distribution to language family size improves with increasing N (for $N = 1 - 5$). This is a kind of behavior familiar from corpus studies of word distributions (Baayen, 2001), where closed-class items – typically function words – yield distributions similar to the 1-grams (phonemes) in this study, whereas open-class words display typical power-law behavior for all corpus sizes, just like the 3–5-grams in this study. We take this as an indication that we are on the right track and investigating a genuine linguistic phenomenon.

Even though this study shows that phoneme N-gram distributions closely mirror the power-law behavior of language family sizes, it raises more questions than it answers about the use of N-gram distributions in linguistic research, such as:

- Is the N-gram distribution an effect strictly connected with genetic relatedness among the languages, or simply an effect of the number of languages in a group (regardless of whether they are related or not)?
- If the effect is genetic, can the size of the family be predicted from N-gram profiles of smaller samples than the full family? (This could be very useful.)
- How well do the N-gram distributions calculated from the ASJP data correlate with phonotactic patterns extracted from larger dictionaries or text corpora, e.g., field notes?

We hope to be able to address these and related questions in the future using the ASJP and other large linguistic databases, as well as corpora of transcribed spoken language, and possibly the output of speech recognition systems.

5. Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Digital areal linguistics*, VR dnr 2009-1448) and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken.

6. References

R. Harald Baayen. 2001. *Word frequency distributions*. Kluwer Academic Publishers, Dordrecht.

Lyle Campbell and William J. Poser. 2008. *Language classification: History and method*. Cambridge University Press.

Harald Hammarström. 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica*, 27(2):197–213.

Paul M. Lewis, editor. 2009. *Ethnologue: Languages of the world*. SIL International, Dallas, TX, USA, Sixteenth edition.

Søren Wichmann, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Sebastian Sauppe, Eric W. Holman, Pamela Brown, Julia Bishoffberger, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, Anthony Grant, and Harald Hammarström. 2011. The ASJP database (version 14). http://email.eva.mpg.de/ wichmann/listss14.zip.

Søren Wichmann. 2005. On the power-law distribution of language family sizes. *Journal of Linguistics*, 41(1):117–131.

Identification of Entities in Swedish

Andreas Salomonsson* **Svetoslav Marinov*** **Pierre Nugues†**

*Findwise AB, Drottninggatan 5
SE-411 14 Gothenburg, Sweden
{andreas.salomonsson,svetoslav.marinov}@findwise.com
†Department of Computer Science, Lund University
S-221 00 Lund, Sweden
pierre.nugues@cs.lth.se

1. Introduction

A crucial aspect of search applications is the possibility to identify named entities in free-form text and provide functionality for entity-based, complex queries towards the indexed data. By enriching each entity with semantically relevant information acquired from outside the text, one can create the foundation for an advanced search application. Thus, given a document about Denmark, where neither of the words *Copenhagen*, *country*, nor *capital* are mentioned, it should be possible to retrieve the document by querying for *capital Copenhagen* or *European country*.

In this paper, we report how we have tackled this problem. We will, however, concentrate only on the two tasks which are central to the solution, namely named entity recognition (NER) and enrichment of the discovered entities by relying on linked data from knowledge bases such as YAGO2 and DBpedia. We remain agnostic to all other details of the search application, which can be implemented in a relatively straight-forward way by using, e.g. Apache Solr[1].

The work deals only with Swedish and is restricted to two domains: news articles[2] and medical texts[3]. As a byproduct, our method achieves state-of-the-art results for Swedish NER and to our knowledge there are no previously published works on employing linked data for Swedish for the two domains at hand.

2. Named Entity Recognition for Swedish

Named entity recognition is an already well-established research area with a number of conferences and evaluations dedicated to the task (e.g. CoNLL 2003 (Tjong et al., 2003)). While many systems have been created for English and other languages, much fewer works have been reported on Swedish.

Dalianis and Åström (2001) and Johannessen et al. (2005) are examples of NER systems for Swedish. Dalianis and Åström (2001) employ rules, lexicons, and machine–learning techniques in order to recognize four types of entities: person, location, organization, and time. It achieves an overall F-score of 0.49 on 100 manually tagged news texts.

The Swedish system described in Johannessen et al. (2005) is rule-based, but relies on shallow parsing and employs gazetteers as well. It can recognize six types of entities: person, organization, location, work, event, and other. It reports an F-score of 0.92 measured on test corpus of 1800 words. On a test set of 40 000 words without gazetteers, the recall of the system drops from 0.91 to 0.53. No information of precision is given by the authors.

3. System Description

3.1 Named Entity Recognition

We have tackled the NER problem by relying entirely on machine–learning techniques. We use a linear classifier, LIBLINEAR (Fan et al., 2008), and train and test the system on the Stockholm-Umeå Corpus 2.0 (SUC) (Ejerhed et al., 2006).

In addition to the POS tags of the tokens, SUC contains information about nine different categories of named entities: person, place, institution, work, animal, product, myth, event, and other. Following the standards in CoNLL 2003, we chose to identify four categories: person, organization, location, and miscellaneous, thus merging the product, myth, event, animal, work and other classes in a miscellaneous category and mapping institution to organization, and place to location.

The most important features of the classifier are: the POS tags of the surrounding and the current token, the word tokens themselves, and the previous two named entity tags. Other features include Booleans to describe the initial capitalization, if the word contains periods, and contains digits. As advocated by Ratinov and Roth (2009), we employ the BILOU (Begin Inside Last Outside Unique) annotation for the entities.

3.2 Linked Data

Linked Data is a part of the semantic web initiative. In its core, it consists of concepts interlinked with each other by RDF triples. This allows us to augment the discovered entities with additional information related to them. We have used the semantic network YAGO2 (Hoffart et al., 2011) and the DBpedia knowledge base. Each named entity we extract from the NER module is mapped to an identifier in YAGO2 if the entity exists in the semantic network. We can then use information about the entity and its relations to other entities. YAGO2 is stored and queried using Sesame RDF repository from openRDF.org, and one can use SPARQL as the query language. One of the most important predicates in YAGO2 for our work is the `isCalled` predicate. Given an entity E we can

[1] `http://lucene/apache.org/solr/`

[2] Articles crawled from dn.se

[3] Articles from 1177.se

Category	Exist	Found	Correct	Precision (%)	Recall (%)	F_1 (%)
Persons	15128	17198	13626	78.17	90.47	83.87
Organizations	6332	4540	3089	68.33	47.49	56.03
Locations	8773	8974	6926	76.97	78.32	77.64
Miscellaneous	3956	2051	1249	64.73	29.60	40.62
Total	34189	32763	24890	75.77	72.35	74.02
Unlabelled	34189	32922	30655	93.11	89.66	91.36

Table 1: NER evaluation

use `SELECT ?id WHERE {?id isCalled "E"}?` to get one or several unique identifiers. If there is only one, we map the entity with its identifier. When multiple identifiers are found, we use a very simple method for disambiguation: We take the identifier with most information related to it.

While YAGO2 was used primarily with the news articles, DBPedia was employed for the medical texts. First each medical term in the SweMESH[4] taxonomy was mapped to an unique identifier from DBPedia. We then searched in the document only for those medical terms which are in SweMESH and use the unique identifier to extract more information about them.

4. Architecture Overview

The core of the system which identifies and augments named entities is implemented as a pipeline. The text of each document passes through the following stages: tokenization, sentence detection, POS tagging, NER, extraction of semantic information. At the end of the pipeline, the augmented document is sent for indexing.

The tokenizer was implemented by modifying the Lucene tokenizer used by Apache Solr. It uses regular expressions to define what a token is. The sentence detector is rule-based and uses information about the types of the tokens, and the tokens themselves as identified in the previous step. The POS tagging stage employs HunPos (Halácsy et al., 2007) which has been trained on SUC. The NER stage is the system which was described in Section 3.1 above. Finally, all named entities are enriched with semantic information if such is available from the knowledge bases.

Given the sentence *Danmark ligger i Nordeuropa.*, the two locations, *Danmark* and *Nordeuropa*, are identified. We then proceed by mapping the two entities to their corresponding identifiers in the semantic network. The YAGO2 identifier for *Danmark* is `http://www.mpii.de/yago/resource/Denmark` and we use it to extract information, e.g. Denmark is a European country; its capital is Copenhagen, with which we then augment the META-data of the document.

5. Results and Discussion

We have evaluated the performance of the NER system with a 10-fold cross-validation on SUC. We used L2-regularized L2-loss support vector classification (primal) as the solver type in LIBLINEAR and all other parameters were set to their default values. While our system achieves very good results for Swedish (see Table 1), it is difficult to compare it to previous systems due to the differences in test data and number of categories. Yet, we see more room for improvement by adding gazetteers and using word clusters as features. In addition, we have noticed inconsistencies in the annotation of entities in SUC. Titles are sometimes part of the entities and sometimes not.

Finally, by enriching entities with related information allows us to retrieve more documents, cluster them in a better way or populate ontologies by using such data. As the extracted information resides in a META field of an indexed document and the field often gets a higher score, the document will get an overall higher rank in the result list.

6. References

Hercules Dalianis and Erik Åström. 2001. SweNam–A Swedish named entity recognizer. Technical report, KTH.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm-umeå corpus version 2.0.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: an open source trigram tagger. In *Proceedings of the ACL 2007 Demo and Poster Sessions*.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011. Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW 2011)*. ACM.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*, 20(1).

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. ACL.

Erik F. Tjong, Kim Sang, and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*.

[4] `http://mesh.kib.ki.se/swemesh/swemesh_se.cfm`

Cohesion in Automatically Created Summaries

Christian Smith, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
christian.smith@liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

In this paper we present results from studies of cohesion in extraction based summaries. Cohesion is measured automatically through the amount of co-references in the text and how intact the text is after summarization. Four different ways of creating 100 word summaries are studied. The texts come from the DUC 2002 corpus of news paper texts. One interesting result is that a summary produced by a traditional vector space based summarizer is not less cohesive than a summary created by taking the most important sentences from the summarizer and the sentence before (at the same time removing the least important sentences), nor has the latter a lower content quality.

1. Introduction

An extraction based summary is created, by one way or the other, extracting the most important sentences from the original text. Cohesion and discourse relations play a vital role in understanding summaries (Louis et al., 2010). The goal is, of course, to create summaries that are readable in that they maintain cohesion while still maintaining a high coverage of content. Previous results have, however, shown that broken or erroneous anaphoric references is a problem in extraction based summarizers (Hassel, 2000; Kaspersson et al., 2012), as they are breaking the cohesion of the summarized text and in some cases even altering the meaning of the text, making them hard for readers to understand.

Pitler et al. (2010) attempted to develop and validate methods for automatic evaluation of linguistic quality in text summarization. Several human judges assessed automatically created summaries with regards to five topics; Grammaticality, Non-redundancy, Referential clarity, Focus and lastly Structure and Coherence. A set of indicators of linguistic quality was developed which then were used to rank summaries according to the five topics. Of these indicators, Continuity-based indicators performed the best in classifying summaries. Continuity-based indicators included Cohesive devices, coreference and cosine similarity.

Of the five topics, the topics of Referential clarity and Structure/Coherence seems to be the most relevant when dealing with extraction based single document summarization since 1) extraction based summarizers seldom have problems with grammaticality (Over et al., 2007), 2) single document summarizer is less likely to repeat information for redundancy and 3) for the same reason as 2) regarding Focus.

In this paper, we present results from studies on cohesion and summary content quality by comparing summaries created using various extraction techniques.

2. Method

In our experiments we created 100 word summaries of the 533 single document news paper texts from DUC 2002.

Four different types of techniques were used to create the summaries:

100FIRST extract the first 100 words, as a baseline, as for news texts the initial paragraphs normally includes the most important part of a text (Nenkova, 2005).

EVERY3 extract every third sentence,

COGSUM use a vector space based extraction summarizer,

PREVSUM use the summarizer and include also the sentence before the highest scored sentences and remove the lowest scored sentences, see Smith and Jönsson (2011).

The summaries were tagged for coreference using the Stanford CoreNLP package[1]. For our investigations we compare the summary to the original based on the following (for reference look at the short example text provided below):

The summer says bye.
It has lasted rather long.
The autumn is nigh.

BROKEN if an anaphor has no antecedent (the first sentence in the example text has been removed).

INTACT if at least one antecedent to an anaphor is extracted (the last sentence in the example is removed).

PARTIAL if the anaphor is missing but the antecedent is in the summary (the second sentence in the example text has been removed).

REMOVED if a coreference chain is completely removed in the summary (The first and second sentences are removed).

Of these, BROKEN and INTACT are the most important as they affect cohesion the most. We also measure summary quality by gold standard comparisons using ROUGE-1.

The summarizer used in our investigations is a Random indexing based summarizer called COGSUM (Smith and Jönsson, 2011). The results however are valid for other vector space based summarization approaches e.g. HolSum (Hassel and Sjöbergh, 2007), Chatterjee and Mohan (2007) and Gong (2001).

[1] nlp.stanford.edu/software/index.shtml

Table 1: Summary quality with regards to content coverage

Summary	Content
EVERY3	0.36558
100FIRST	0.45925
COGSUM	0.39942
PREVSUM	0.38613

3. Results

Figure 1 depicts the results form our cohesion studies and Table 1 shows how the different summarizers performed.

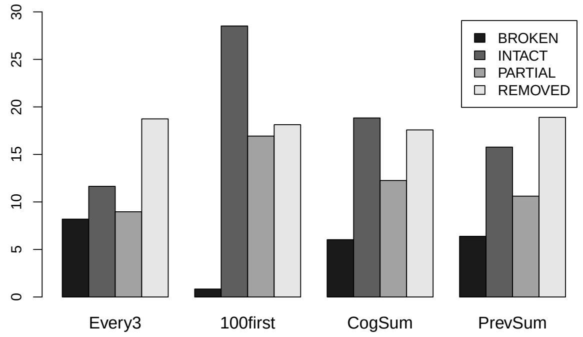


Figure 1: Cohesion measures on the four different summarizers.

Considering Content, 100FIRST is significantly better than all the other ($p < .05$). COGSUM and PREVSUM are significantly better than EVERY3 ($p < .05$).

Comparing the cohesion there are a variety of significances, for instance, for broken references the 100FIRST is significantly better than all the other ($p < .001$).

4. Conclusion

As expected the first 100 words of a text gives the best summary both in terms of cohesion but also summary content. This is true for short newspaper articles but probably not for longer texts from other genres. COGSUM, in its two versions, produce better summaries than taking every third sentence.

Taking the previous sentences in COGSUM doesn't affect the content quality and only slightly increases cohesion. This is interesting, as there is always a tradeoff between the amount of new information to include in the summary and the cohesion of the text. If some important sentences are disregarded and instead sentences that improve cohesion are included there is a risk that the summary will be less informative, but that was not the case in our studies using news texts.

Using news texts has its limitations, as also pointed out by Over et al. (2007), but this is where most current research is conducted, and is, thus, an interesting starting point. Further experiments with other text types may show that summarizers that consider cohesion give more readable summaries without significantly losing content.

5. References

Nilhadri Chatterjee and Shiwali Mohan. 2007. Extraction-based single-document summarization using random indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

M Franzke, E Kintsch, D Caccamise, N Johnson, and S Dooley. 2005. Summary street®: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33(1):53–80.

Yihong Gong. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Martin Hassel and Jonas Sjöbergh. 2007. Widening the holsum search scope. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida)*, Tartu, Estonia, May.

Martin Hassel. 2000. Pronominal resolution in automatic text summarisation. Master's thesis, Master thesis in Computer Science, Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden.

Thomas Kaspersson, Christian Smith, Henrik Danielsson, and Arne Jönsson. 2012. This also affects the context - errors in extraction based summaries. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.*

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Tokyo, Japan*, pages 147–156.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1436–1441. AAAI Press / The MIT Press.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43:1506–1520, Jan.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality inmulti-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 544–554.

Christian Smith and Arne Jönsson. 2011. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand.*

Statistical Identification of Pleonastic Pronouns

Marcus Stamborg **Pierre Nugues**

Lunds University, Department of Computer Science
Lund, Sweden
cid03mst@student.lu.se, Pierre.Nugues@cs.lth.se

Abstract

This paper describes an algorithm to identify pleonastic pronouns using statistical techniques. The training step uses a coreference annotated corpus of English and focuses on a set of pronouns such as *it*. As far as we know, there is no corpus with a pleonastic annotation. The main idea of the algorithm was then to recast the definition of pleonastic pronouns as pronouns that never occur in a coreference chain. We integrated this algorithm in an existing coreference solver (Björkelund and Nugues, 2011) and we measured the overall performance gains brought by the pleonastic *it* removal. We observed an improvement of 0.42 from 59.15 of the CoNLL score. The complete system (Stamborg et al., 2012) participated in the CoNLL 2012 shared task (Pradhan et al., 2012), where it obtained the 4th rank.

1. Introduction

In this paper, we describe a method to identify pleonastic pronouns. Our work was motivated by a participation in the CoNLL 2012 evaluation on coreference solving in English, Arabic, and Chinese (Pradhan et al., 2012). Popular statistical algorithms to solve coreference such as Soon et al. (2001) use a two-step procedure, where they first extract candidate mentions, usually all the noun phrases and pronouns, and then apply a classifier to pairs of mentions to decide whether they corefer or not.

The mention extraction stage was originally designed to reach a high recall, i.e. build a large set of mentions from which the coreference chains are extracted. A consequence of this lack of selection is that it creates a large number of false positives. Starting from a coreference solver by Björkelund and Nugues (2011), that does not include a pleonastic pronoun identification, we could observe that the pronoun *it* stood out with the worst performance.

We designed a preprocessing stage to identify automatically the pleonastic pronouns and remove them from the set of mentions before they are passed to the classifier. We added this stage to the coreference solver and we report here the improvements we obtained.

2. Previous Work

The idea to remove pleonastic *it*s has been used in a couple of coreference solvers. An example is the high-performance Stanford solver (Lee et al., 2011) that includes a simple rule-based module to identify these pronouns. The rules consider the current word and the word following in the sentence. If the current word is *it* and any of the following words:

> *is*, *was*, *seems*, *seemed*, *appears*, *looks*, *means*,
> *follows*, *turns*, *turned*, *become*, *became*,

are found immediately after it, it is tagged as pleonastic and discarded from the mention list.

3. Classifier

To design a classifier, we used the approximation that noncoreferring pronouns i.e. pronouns not member of a coreference chain in the annotated corpus were pleonastic. By definition, pleonastic pronouns are outside coreference chains. However, using this idea we fail to identify singleton pronouns that lack antecedents.

We trained a classifier using logistic regression and the LIBLINEAR package (Fan et al., 2008). The training data was selected by extracting all the instances of the word *it* from the corpus. We used a small pool of features that we selected with a simple greedy forward/backward selection. Table 1, left column, shows the initial feature set that was selected at this stage in the development. We applied this pleonastic detector as a preprocessing step and using the complete original coreference solver, we could observe a slight increase of the score.

3.1 Pre/Postprocessing

In the initial trials, we used a preprocessor to remove the pleonastic pronouns from the mentions. We also tried to move the removal as a postprocessing stage, where we discarded the pronouns from the coreference chains. Although giving an increase of the overall score, it was lower than by using the preprocessor and we did not follow this path.

3.2 Combination of Probabilities

We noticed that isolated pleonastic identifier modules, either as pre or postprocessors, removed a significant portion of nonpleonastic *it*s. We introduced a second term in the classifier to take into account the likelihood that the pronoun was part of a coreferring pair. We used the probability that the word was pleonastic, P_{pleo}, together with the result from the coreference resolver, P_{coref}. We applied the inequality:

$$P_{coref}(\text{Antecedent}, it) \times (1 - P_{pleo}(it)) > 0.4,$$

to decide on the pleonastic nature of *it*.

The only change when the word *it* is one of the mentions is that the ordinary output from the coreference classifier is scaled by the pleonastic classifier. We found the cutoff value of 0.4 experimentally, using 5-fold cross-validation.

Using the probability combination, we carried out a second feature selection and Table 1 shows the final feature set, right column.

Initial set	Final set
HeadLex	HeadLex
NextWordLex	HeadRightSiblingPOS
—	HeadPOS

Table 1: The feature set used by the pleonastic *it* classifier.

English 2011	CoNLL score
Baseline	53.27
Handwritten rules	53.21
Pre-processor	53.51
Post-processor	53.63
Combination of probabilities	**53.90**

Table 2: Scores on the 2011 English development set (Pradhan et al., 2011) using various ways of removing pleonastic *it* pronouns. The rules were based on those used by the Stanford coreference solver (Lee et al., 2011).

3.3 Results

We carried out the initial testing with the 2011 CoNLL shared task corpus (Pradhan et al., 2011) and the original coreference system by Björkelund and Nugues (2011). Table 2 shows the results for the various alternatives we tested.

The coreference system we submitted to CoNLL 2012 is significantly different, notably because it handles multiple languages and it uses a different feature set for the identification of coreferring pairs. Table 3 shows the final scores we obtained on the development set with and without the pleonastic identification using the CoNLL 2012 system. We obtained similar increases using a cross-validation. We report the results on the CoNLL 2012 corpus which are slightly different from those of the CoNLL 2011 corpus.

In the development set, there are 1402 occurrences of the *it* pronoun; out of them 792 are part of a coreference chain and 610 are not. Table 4 shows the amount of *it* tagged either as coreferring or not in the output file created by the system. As can be seen in Table 4, the number of false negatives increased while the number of false positives decreased when applying a pleonastic detection.

4. Conclusions

In order to obtain good results during coreference resolving, it is important to have a high recall, but increasing the precision has proven beneficial as well as demonstrated by the pleonastic *it* addition.

Despite a relatively larger number of false negatives, we observed an increase to the overall score, which indicates that it is better to remove as many false positives as possible despite increasing the number of false negatives. Increasing the accuracy of the pleonastic *it* classifier, for example by crafting more, possibly better features would likely lower the amount of false negatives while increasing or at least retaining the amount of false positives.

English 2012	CoNLL score
Without removal	59.15
With removal	**59.57**

Table 3: Scores on the 2012 English development set (Pradhan et al., 2012) with and without removal of the pleonastic *it* pronouns.

Set	No pleo. module	Pleo. module
Coreferring	1318	966
Noncoreferring	84	436
False Positives:	556	327
False Negatives:	30	153

Table 4: Counts of *it* classified as part of a chain (coreferring) or not (noncoreferring) by the coreference system, with and without the pleonastic *it* module. The positive set is the set of coreferring it pronouns.

Acknowledgments

This research was supported by Vetenskapsrådet, the Swedish research council, under grant 621-2010-4800, and the European Union's seventh framework program (FP7/2007-2013) under grant agreement no. 230902.

5. References

Anders Björkelund and Pierre Nugues. 2011. Exploring lexicalized features for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 45–50, Portland, Oregon, USA, June.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, June.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Marcus Stamborg, Dennis Medved, Peter Exner, and Pierre Nugues. 2012. Using syntactic dependencies to solve coreferences. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 64–70, Jeju Island, Korea, July.

Telling questions from statements in spoken dialogue systems

Sofia Strömbergsson, David House, Jens Edlund
KTH Speech, Musik and Hearing
Stockholm, Sweden
sostr@csc.kth.se, {edlund,davidh}@speech.kth.se

1. Domain dependency: controlling context, style and situation

To date, just about every system that successfully communicates with humans using language owes a great deal of its success to one particular set of characteristics. The ability to control or make assumptions about the linguistic context, the dialogue style and the physical situation in which the dialogue occurs is essential to systems ranging from early text based systems such as Eliza (Weizenbaum, 1966) to modern speech based question-answer systems such as Siri, acquired by Apple in 2010 and launched the companies' smart phones in 2011. Controlling or predicting context, style and situation allows us to build *domain dependent* systems – systems designed to handle a very small subset of the dialogues humans in which take part.

One of the motivations for spoken dialogue system research is an improved understanding of human interaction. But demonstrating an understanding of human interaction by mimicking it in a machine requires that we do not constrain the domain any more than humans do. A widely held hope is that domain independency can be reached by (1) gradually increasing in-domain coverage, (2) gradually widening the domains, (3) adding new domains, (4) developing robust domain detection, and (5) methods to seamlessly move from one domain to another. The soothing idea is that progress can be gradual, and that each small improvement adds to the improvement of the whole.

When subjected to closer scrutiny, the hope of gradually building ourselves out of domain dependency may be overly optimistic. Some distinctions that seem trivial to humans and spoken dialogue systems alike may turn out to be unexpectedly difficult when we relinquish control over context, style and situation. We suspect that the simple distinction between question and statement is an example of this.

2. Dialogue acts: questions and statements

Deciding whether an utterance is a question or a statement is relatively uncontroversial for humans. In some cases, questions are clearly marked lexically, syntactically, or both: wh-questions have an initial wh-word, and yes/no questions have their tell-tale word order. In other cases, utterances that are pragmatic questions – utterances that function as a question – lack these distinctive features and utterances that are superficially questions do not have question function. Yet people rarely find their understanding problematic. The following dialogue snippet from a fictitious interview in the 2000 film *How to kill your neighbor's dog* is an exception:

Debra: There's no denying that in the 80s you were quite the boy wonder. All your plays went on to Broadway. That hasn't been the case in the 90s, you've had something like three bombs in a row. [SILENCE]. Peter?

Peter: I didn't hear a question in there.

Statements like Peter's are common in political debate, citations from which make up a substantial part of the 100000+ hits Google delivers for "I did not hear/detect a question". Outside of rhetoric and comedy, however, it is difficult to find examples of people mistaking questions for statements and vice versa.

3. Questions and spoken dialogue systems

As with people, telling questions and statements apart is not on any list of hard problems for current spoken dialogue systems. This, however, is largely because the systems are designed so that speech acts are predictable. There is a correlation between initiative and question, so that a dialogues with system initiative will rarely allow questions from the user, so user questions will not be understood and the non-understanding will be explained as out-of-domain. In the case of user-initiative dialogue, the opposite is true, and most user utterances are expected to be questions. Even mixed-initiative dialogues often constrain the users' freedom. In the mixed-initiative apartment browsing system Adapt (Gustafson et al., 2000), for example, a dialogue consisted of two phases. In the first phase, the system would ask the user to provide constraints to narrow down the search. When the number of suitable apartments was 7 or less, the system entered the second phase, in which the user asks questions about the apartments, and the system answers.

These tricks are lost in a move towards humanlike systems engaging in everyday conversations with people, and the question/statement distinction becomes less trivial for spoken dialogue systems. For this reason, we investigate 600 questions and question-like utterances from the Spontal corpus of conversational Swedish with a view to cues as to how a spoken dialogue system can best tell them apart from statements.

4. Hypothesis

This work is preliminary, but it takes as its starting point the analyses we have already done on questions in the Spontal corpus, which suggest in particular that (1) the concept of question is difficult to define and (2) question markers are not always present. Dialogue act classifications are often quite circular when it comes to defining questions. As an example, speech act classification used in the Edinburgh Map Task (Anderson et al., 1991), for uses "question" both in label and description and hedges with a catch-all definition of wh-questions: "Any query not covered by the other categories". We expect to find, then, that the identification of questions is far from easy, but rather is one of the key problems to solve for spoken dialogue systems to take a leap towards achieving human conversational skills.

5. Material

Questions and matching non-questions in this study were extracted from orthographically transcribed subset of 24 dialogues from the Spontal corpus (Edlund et al., 2010). Two annotators looked for questions while transcribing the spoken dialogues and labeled these with a question tag. The definition of "question" was deliberately kept quite open: "Anything that resembles, structurally or functionally, in whole or in part, a question". In all, 908 talkspurts received the question label.

Three independent annotators labeled all 908 instances with respect to four relatively simple queries. (For a detailed description of the queries, see Strömbergsson et al., in press-a). During this process, annotators could choose to *skip* talkspurts that they felt were in no sense a question, or that were otherwise impossible to judge. Talkspurts that were skipped by at least one annotator, or were all annotators had disagreed on the question label, were excluded from further analysis, leaving a set of 641 questions. The targeted 600 questions were selected from this set so that they were balanced for the interlocutors' gender and previous acquaintance.

For each of the 600 questions, a matching non-question was selected from the same speaker and the same dialogue as the question, with the restriction that a) the difference in duration between the question and the non-question, and b) the time between the question and the non-question in the dialogue, were minimized. (Talkspurts that had been labeled with a question tag earlier were not eligible for selection.)

6. Prosodic measures

Five different prosodic measures were extracted from the questions and non-questions: speech rate (RATE), average pitch (AVG), pitch variation (VAR), pitch range (RNG), and an estimate of intonation slope (DIFF). (For descriptions of these measures, see Strömbergsson et al., in press-b).

7. Results

In order to explore prosodic (within-speaker) differences between questions and matching non-questions, multiple t-tests were conducted. Table 1 shows the results of these tests.

Measure		**Descriptives**	**t-test statistics**

RATE	Qns
	nQns
		M = 5.2, SD = 1.9
		M = 3.84, SD = 2.11
			t(599) = 12.0, p < .001

AVG	Qns
	nQns
		M = 19.9, SD = 4.6
		M = 19.5, SD = 4.4
			t(599) = 3.7, p < .001

VAR	Qns nQns	M = 5.8, SD = 1.3 M = 5.8, SD = 1.3	t(599) = -,8, p = .42

RNG	Qns
	nQns
		M = 16.3, SD = 6.3
		M = 16.5, SD = 6.2
			t(599) = -1.0, p = .32

DIFF	Qns
	nQns
		M = -.3, SD = 2.6
		M = -.3, SD = 2.9
			t(599) = -.5, p = .60

Table 1: Descriptive and t-test statistics for the five prosodic measures, for questions and non-questions.

The figures presented in Table 1 suggest that questions are produced faster than non-questions, and with slightly higher pitch. No other differences are found.

The finding that questions are produced at a faster speaking rate than non-questions is surprising and could be explained by the fact that the questions have all been selected so that they contain nothing more than the actual question, whereas we do not have the same control for the non-questions. The non-questions can (and do) also contain other vocalizations, such as laughter and breathing, which contributes to longer duration but without an accompanying increase in the number of consonant-to-vowel transitions. Thus, speaking rate appears to be slower.

8. Discussion and future work

The two-page abstract format for SLTC publications does not provide room for further detail, nor for discussion. We are forced refer the reader to the accompanying poster presentation and to forthcoming full publications.

9. Acknowledgements

This work is funded by the Swedish Research Council, Humanities and Social Sciences (VR 2009-1764) "Intonational variation in questions in Swedish".

10. Reference

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992 - 2995). Valetta, Malta.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., & Wirén, M. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proc. of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 134-137). Beijing: China Military Friendship Publish.

Strömbergsson, S., Edlund, J., & House, D. (2012a). Questions and reported speech in Swedish dialogues. In *Proc. of Nordic Prosody XI*. Tartu, Estonia.

Strömbergsson, S., Edlund, J., & House, D. (in press-b). Prosodic measurements and question types in the Spontal corpus of Swedish dialogues. To be published in *Proc. of Interspeech 2012*. Portland, Oregon, US.

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery, 9*, 36-45.

On the Interplay between Readability, Summarization, and MTranslatability

Sara Stymne[1,2], **Christian Smith**[1,3]

[1]Linköping University, [2]Uppsala University, [3]Santa Anna IT Research Institute AB
sara.stymne@lingfil.uu.se, christian.smith@liu.se

1. Introduction

NLP can be used in several ways to help people read texts that are inaccessible to them in some way. Machine translation can be used to translate texts from a language one do not know, and summarization can be used to shorten texts in order to make them faster and/or easier to read. Both these techniques do in a sense improve readability. In this paper we describe a small initial study of the interplay between readability, summarization, and MTranslatability. We applied an extractive summarizer to a set of news stories, and investigated the effect on readability and machine translation (MT).

Extractive summarizations are created by extracting the most important sentences from a text; no re-writing of sentences takes place. Smith and Jönsson (2011a) showed that by using extractive summarization, readability was improved according to several readability measures. Their study was performed on Swedish text, and showed an effect on several genres including news.

MTranslatability is the translatability of a text by an MT system, i.e., how easy a text is for an MT system to translate (Bernth and Gdaniec, 2001). This notion has mostly been exploited in connection to rule-based MT systems, for instance by improving MT output through re-writing the source according to some controlled language rule set (see e.g Roturier (2004)). MTranslatability is also related to the notion of confidence estimation (CE), to estimate the quality of machine translated sentences (Specia et al., 2009). Many features used for CE are also similar to those used in many readability formulas, such as sentence length. CE differs from the notion of MTranslatability, however, in that in CE the estimation is done after the translation phase, whereas MTranslatability is related to assessing whether a sentence or text is easy or hard to translate before it is sent to the MT system.

2. Experiment

We performed an experiment where we applied an automatic summarizer to English news text, which was then translated by a statistical machine translation system into German. We then investigated how the summarization affected readability and MT.

2.1 Summarizer

We used the extractive summarizer CogSum (Jönsson et al., 2010). It is based on the vector space model random indexing (RI). Each sentence in a document is scored by RI, and then ranked using a weighted PageRank algorithm. A summary is then created by choosing the top sentences, which supposedly are the most important to the document. The length of the summaries can be varied; in the current study we use 50% summaries, i.e., extract half the sentences from each text.

We used a version of CogSum with an external corpus, the Brown Corpus, for training the vector space, which has been shown to be more stable than training the vector space on the document to be summarized (Smith and Jönsson, 2011b).

2.2 Readability measures

Readability is a complex notion, that is related both to texts, and to individual readers and their skills. However, many automatic readability measures have previously been proposed based on, often superficial, text properties. In the current study we chose to work with two such measures: the Flesch reading ease formula (Flesch, 1951), which has commonly been used for English, and LIX (Björnsson, 1968), which has traditionally been used for Swedish. Flesch is based on average sentence length (ASL) and average syllable length, and LIX is based on ASL and the proportion of long words. In addition we report ASL on its own.

2.3 Machine translation system

We used a factored phrase-based statistical MT system (Koehn and Hoang, 2007) with treatment of compound words, to translate from English to German, as described in Holmqvist et al. (2011). The system was trained on 1.7M sentences from Europarl, 136K sentences of News text, and an additional 17M sentences of monolingual news data. The system is evaluated using the Bleu metric (Papineni et al., 2002).

2.4 Data set

The experiment was performed on news stories from the development data of the WMT12 workshop.[1] From this data we extracted all coherent news stories with at least 20 sentences, resulting in 236 news stories, with a total of 7602 sentences. The stories were translated by the SMT system, and an analysis was performed on both the full set of sentences, and on sentences that occurred in the summaries, compared to sentences not in the summaries. As we found that the summarization affected the average sentence length in these sets, we also used test sets balanced for sentence length, by extracting an equal amount of sentences in all three conditions, with lengths between 13–30 words per sentence, in total 1556 sentences.

[1] http://statmt.org/wmt12

3. Result

In this section we first present the results of summarization on readability, followed by the effect of summarization and readability indicators on machine translation.

3.1 Readability for summaries

To analyze the effect on readability of the summarization we computed readability measures for each of the 236 stories on the full test set (*All*), the summarized sentences (*Sum*), and the sentences not chosen for summarization (*Nosum*). A summary of the results is shown in Table 1. For Flesch, a high score means more readable, whereas a low score is better for LIX and ASL. All differences from the original texts are significant with $p < 0.01$. On all three measures the summaries are less readable than the original texts. We also found that there is a strong significant correlation of -0.91 between Flesch and LIX, something that we believe has not been shown before, since these metrics have tended to be used for different languages.

	Flesch↑	LIX↓	ASL↓
All	44.1	49.6	21.4
Sum	39.6	54.1	27.0
Nosum	47.9	45.6	15.6

Table 1: Readability before and after summarization

3.2 Effects on MTranslatability

Table 2 shows the Bleu scores for the three data partitions, both full and normalized in length. Note that the test sets are different, for the full sets they even have different size, and are thus not directly comparable. The scores, however, indicate that the summaries, which were the least readable, have lower scores than the non-summaries, even when length is compensated for.

	Full sets	length normalized
All	15.4	15.0
Sum	15.1	15.2
Nosum	15.7	15.7

Table 2: Bleu scores

To allow a further analysis of the interaction between SMT, readability and summarization, we performed a multiple linear regression analysis on blocks of 50 sentences for each test set. We could find no significant correlations between Bleu and readability measures or summarization set. We could, however, confirm the correlation between Flesch and LIX, which in this configuration was -.92. Figure 1 illustrates this for the full test sets.

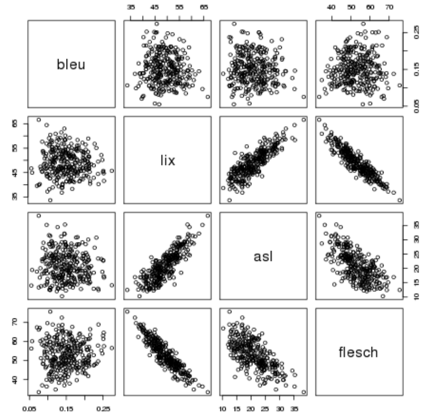


Figure 1: Plot of Bleu against readability metrics

4. Conclusion

In this study summarization resulted in longer and less readable sentences, which is contrary to previous research (Smith and Jönsson, 2011a). The language used and summarizer configuration were different in these studies, however. We plan to conduct further research on the effect of different summarization methods in future work.

None of the readability factors or summarization sets contributed to any significant differences in MT quality in this study. The overall Bleu scores, though, at least seemed to indicate that there is some relation to readability, since the nosum test set had both the highest Bleu scores, and the best readability. The current study has several limitations, though, such as the use of automtic metrics. For instance, even though Bleu has been shown to have some correlation with human judgment, it might not be the best option for this type of study, as in CE, where human judgments have been more useful than metrics. We believe that these issues are interesting, and that further analysis of the interplay between summarization, MTranslatability, and readability could be valuable.

5. References

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16:175–218.

C. H. Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Rudolf Franz Flesch. 1951. *How to test readability*. Harper, New York.

Maria Holmqvist, Sara Stymne, and Lars Ahrenberg. 2011. Experiments with word alignment, normalization and clause reordering for SMT between English and German. In *WMT'11*.

Arne Jönsson et al. 2010. Skim reading of audio information. In *SLTC'10*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP/CoNLL'07*, pages 868–876, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL'02*.

Johann Roturier. 2004. Assessing the set of controlled language rules: can they improve the performance of commercial machine translation systems? In *Translating and the Computer 26*.

Christian Smith and Arne Jönsson. 2011a. Automatic summarization as means of simplifying texts, an evaluation for Swedish. In *Nodalida'2010*.

Christian Smith and Arne Jönsson. 2011b. Enhancing extraction based summarization with outside word space. In *IJCNLP'11*.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *EAMT'2009*.

Opening OPUS for User Contributions

Jörg Tiedemann, Matthias Zumpe, Sebastian Schleussner

Department of Linguistics and Philology
Uppsala University
{firstname.lastname}@lingfil.uu.se

Abstract

OPUS is a growing collection of freely available parallel corpora. In this paper, we describe our on-going effort in developing a web application that opens the resource repository for external users who wish to contribute to the data collection. The system, we describe, implements a flexible, distributed repository backend that includes tools for automatic import of documents in a variety of formats and other features supporting data management and the creation of parallel aligned corpora. Currently, we work on a web frontend that implements the user interface to the repository software.

1. Introduction

OPUS (http://opus.lingfil.uu.se) is a widely used data collection of parallel corpora, tools and on-line corpus search interfaces. It currently covers over 90 languages and more than 40 billion tokens in 2.7 billion parallel translation units (Tiedemann, 2012). It is a growing resource mainly used by people working in data-driven machine translation (for example, statistical MT), cross-lingual corpus linguistics and translation studies. The demand of parallel corpora and related tools is high which can be seen in our web logs showing between 20,000 and 30,000 unique visitors every month. Our goal for the near future is to release a system that makes it easier to contribute to our collection. This system is designed to support all the major steps that need to be done when creating parallel aligned corpora from translated resources. The main purpose is to provide a simple interface to various tools that can be used to validate, convert and align documents contributed by external users in a variety of popular formats and to automatically create parallel data sets in the internal OPUS format. Such a system needs to be scalable and modular in order to cope with the requirements of a multi-user web application. Related initiatives collecting and integrating user-provided translation data exist but usually belong to commercial companies like Google (http://translate.google.com/toolkit) and MyMemory (http://mymemory.translated.net) or pay-services like the one provided by TAUS (http://www.tausdata.org). The OPUS repository software, on the other hand, is an open-source project and data sets will be collected to increase the amount of freely available parallel corpora within OPUS.

In the following, we briefly describe the repository backend, the data import features and the on-going work on a web frontend for OPUS.

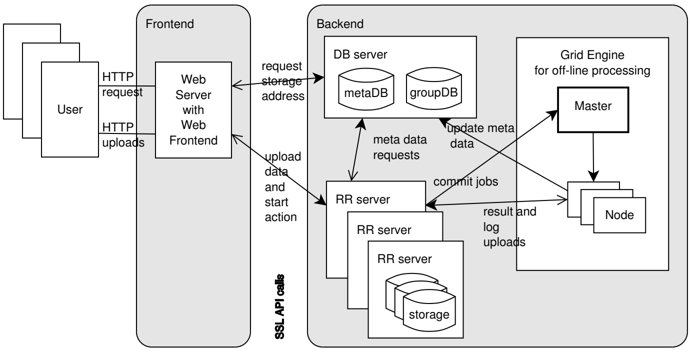


Figure 1: The general architecture of the system.

2. The Resource Repository Backend

The resource repository is a highly modular toolbox that is designed to be flexible and scalable. Figure 1 illustrates the general architecture of the software. The communication with the web frontend is handled via HTTPS requests. The package provides several web service API's to perform tasks at the backend and to obtain information from the resource collection. The backend system includes a central database server (DB server). Its task is to manage data permissions, group settings and the entire collection of meta information available for each resource in the repository. Each of the repository servers (RR servers) is connected to that database via a client-server architecture. For our purposes, we use a schema less key-value store, which makes it possible to easily extend and adjust the database without changing settings or internal data structures. The DB engine is based on TokyoCabinet (FAL, 2010), a modern implementation of a database management system that supports the storage of arbitrary data records without predefined typed data structures. Using the table extension of the software, we can attach records of key-value pairs to each resource identifier (which we use as unique table key). The query engine supported by TokyoCabinet is very powerful and fits our needs very well. Various search operators can be used to access key-value pairs and access is extremely fast. For remote access to the database server, we use TokyoTyrant, a network interface to TokyoCabinet. This server runs as a multi-threaded daemon on the DB server and TokyoTyrant clients connect from other RR servers via dedicated network ports.

Furthermore, RR servers are also configured as submit hosts of a general purpose grid engine. We use the Oracle Grid Engine (formerly Sun Grid Engine, SGE), a standard open-source implementation for Grid Computing, to perform tasks such as data conversion and alignment. In this way, we can easily distribute jobs among several execution hosts, which makes the system highly scalable and extensible. For additional scalability, it is also possible to distribute resources over several storage servers. A load balancer at the frontend could be used to automate this task. However, such a component is not implemented yet.

Using our modular design, the system supports highly distributed setups. Resource repositories may be placed on local networks or even in a cloud-based solution. The database also supports replication on remote servers. RR servers and SGE execution hosts can be added on demand.

2.1 Data Upload

The resource repository supports the upload of documents in various formats. Each data upload triggers an automatic import process that includes a validation step, a conversion step and possibly another sentence alignment step in case parallel documents can be identified.

The software currently supports the following upload formats. **Parallel data sets:** Documents may already be aligned on some segmentation level (usually sentences). We support the three most common formats used in the translation business, the localization industry and in the machine translation community.

- TMX: Translation memories in translation memory exchange format (TMX). Character encoding is automatically identified and language specifications are taken from the language attributes of the translation units.
- XLIFF: Another translation memory format which is primarily used by the localization industry.
- Moses format: Archives of plain text files that are aligned on the sentence level. Languages are taken from the file extension as commonly used in the Moses SMT platform.(Koehn et al., 2007)

TMX and XLIFF formats are checked with validating XML parsers. Minor problems are solved automatically using XML cleaning tools.

Monolingual documents can be in PDF format, various MS Word formats (doc and docx) and in plain text with various character encodings. Recently, we also added the support of the movie subtitle format SRT, HTML, generic XML and derived formats, the Open Document Format and RTF. The software uses standard tools such as Apache Tika (Mattmann and Zitting, 2011) for the conversion to text and, finally, our internal XML format. Character encodings are detected automatically using language-specific classifiers, byte-order markers and file content heuristics.

Each document added to a corpus in the repository is also checked by a language identifier (based on textcat (van Noord, 2010)). Mismatches with user specifications are marked in the metadata DB. The software supports also different settings of the automatic import. Import parameters can be controlled via user-specific, corpus-specific and resource-specific parameters. The same applies for the automatic sentence alignment. The system looks for parallel documents in the repository each time a new upload has been processed. In the default mode, only documents with identical names will be detected as parallel documents. However, the system also supports a fuzzy matching approach using name heuristics and file size ratios. The software also integrated several alignment tools (Gale and Church, 1993; Varga et al., 2005) and supports various settings of the underlying algorithms. Import and alignment processes can be restarted using different parameters using the Job API of the package.

3. The OPUS Web Interface

The task for the web interface is to provide an intuitive graphical system that allows external users to use the repository software and to inspect the data collections provided by its users. The interface is developed using a modern web framework called *mojolicious* (http://mojolicio.us/). This framework is a powerful toolbox supporting modern web technology including many plugins and extensions. The web interface supports all main features of the repository including data upload, download, inspection and other management tasks. It also implements a simple user management and is ready to be tested in a production environment.

4. Conclusions

In this paper we present the main components of a newly developed resource repository that will be used for the creation of parallel corpora from various sources. The system uses modern web technology with a distributed modular architecture in order to support external users when building new resources and contributing to OPUS.

5. References

FAL Labs, 2010. *Fundamental Specifications of Tokyo Cabinet.* The software is available from `http://fallabs.com/tokyocabinet/`.

William A. Gale and Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Chris A. Mattmann and Jukka L. Zitting. 2011. *TIKA in Action.* Manning Publications Co. Apache Tika is available at http://tika.apache.org/.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Gertjan van Noord. 2010. TextCat - an implementation of a text categorization algorithm. Software available at `http://www.let.rug.nl/~vannoord/TextCat/`.

D. Varga, L. Nmeth, P. Halcsy, A. Kornai, V. Trn, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

An audience response system-based approach to speech synthesis evaluation

Christina Tånnander
The Swedish library of talking books and Braille (TPB)
Box 5113, 121 17 Johanneshov, Sweden
christina.tannander@tpb.se

Abstract

Traditional evaluation methods for speech synthesis typically involve tests of very short samples, in which a minimal fraction of the synthesis is evaluated. The need for tests of larger samples, where subjects listen to longer speech chunks, is met by using a method inspired by Audience Response Systems, where subjects are asked to click whenever they hear, for example, something unintelligible. With this test method, it is possible to quickly and efficiently evaluate synthetic voices in a way that better reflects the situation of our end users: people with print impairments listening to long texts such as text books and newspapers.

1. Introduction

TPB is a governmental body providing people with print impairments with Braille and talking books. Our services include university level text books and newspaper produced with speech synthesis, which makes the question of speech synthesis evaluation central for reaching our aim: to produce accessible literature quickly while maintaining the best quality. This goal cannot be met without efficient and reliable speech synthesis evaluation methods.

Traditional evaluation methods of speech synthesis involve tests of brief samples, where the subjects listen to short chunks of synthesized speech and grade or compare them in one way or another. They can be asked to judge for example whether the speech is "intelligible", "natural" or if it is a "pleasant voice". Commonly used evaluation methods include:

- *Grading tests*, which are commonly used to evaluate a certain voice or to compare different voices, and have been used in combination with other test methods in the Blizzard Challenge evaluation (Black & Tokuda, 2005). The subjects can be asked to give a general judgment, or to grade a specific feature such as intelligibility or naturalness.
- *Discrimination tests*, which are brief-sample tests, where the subjects are asked to repeat or write down what they hear, or simply choose between two or more given words. A variant of discrimination tests is to use semantically unpredictable sentences (SUS) (e.g. Benoit, 1989).
- *Preference tests*, which provide an easier task for the subjects: they compare two or more utterances and decide which of them they prefer.
- *Comprehension tests*, which differ from the above-mentioned test methods, as they do not involve a judgment task by the subjects; they listen to a text and are asked to answer questions about what they heard.

For a more thorough compilation of evaluation methods with a different organization, see for example Heuven and Bezooijen (1995).

Most of these methods tests brief sound files of a word or perhaps a couple of sentences, which makes them time-consuming for tests of realistic amounts of data (from a text book or newspaper perspective). The first three methods also invite over-analysing, which again increases the time subjects put into the task. Furthermore, it is crucial but far from straight-forward to formulate the right questions to ask, to make sure that the subjects understand them, and to understand the subjects' answers. Edlund (2011) describes the problem as follows:

> *[...]a rather large body of research suggests that participants cannot reliably account for their perception. Although perhaps the strongest lesson to be learned from this is that it is preferable to measure the effects of stimuli and events in subjects rather than asking the subjects how they perceive them [...]* (pp.204-5)

2. Audience response system

For our purposes, we need a method with higher ecological validity - one that better reflects the needs of the real end user of our speech syntheses: people with print impairments listening to long texts. For this reason, we are experimenting with tests inspired by Audience Response Systems (ARS). ARS allow concurrent evaluations by many subjects, simply and time-efficiently, by asking subjects to click whenever they like (or dislike) something while watching or listening to a continuous stimuli. The method has been used for evaluations of movies (i.e. screenings), to increase engagement in the classroom, and for voting in TV shows and the like. A more detailed background is given in Edlund (2011).

For speech synthesis evaluation, we may ask subjects to click when they hear something they don't like, a certain word or a specific feature, when they think the speech is too fast or don't understand what is said. ARS also allows us to give relatively vague instructions, rather than specify a number of more or less dense features to judge, such as naturalness or intelligibility.

3. Method

24 subjects were divided into three groups with eight subjects in each. Parameters such as gender, age, speech synthesis experience or print impairments were not considered.

The test tool used is developed by STTS (Södermalms Talteknologiservice AB), and is a web-based tool for grading tests, preference tests and ARS inspired click tests.

The test material consists of four such sound files from different types of university level text books, to reflect the normal domain of the synthesis. In order to make the subjects familiar with the test situation, the test started with a practice sound file of 28 seconds. This file is not part of the results. All subjects listened to the same sound files in the same order.

The three test groups were given the following instructions:

Group 1 click when you hear something unintelligible

Group2 click when you hear something irritating

Group 3 click when you hear something not entirely correct

4. Results

The results show that there is a clear difference between the number of clicks of group 1 and the other two groups:

Group	Total number of clicks	Mean number of clicks per file and subject
1	28	0.88
2	272	8.50
3	294	9.18

Table 1. Total number of clicks per group and mean number of clicks per file and subject.

The individual number of total clicks shows that there is little variation in group 1, while the subjects’ clicks in group 2 vary between 8 and 76, and in group 3 between 2 and 72. The standard deviation for group 1 is 2.06, for group 2 22.19 and for group 3 22.79.

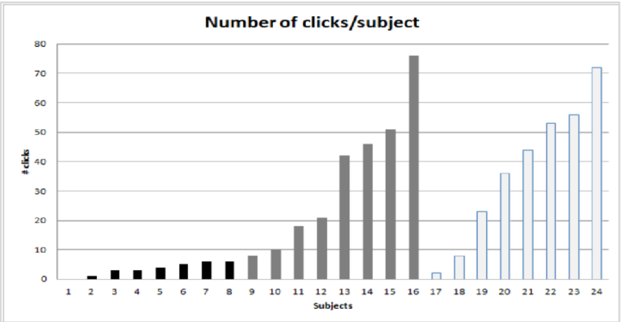


Figure 1. Number of clicks/subject arranged from lowest to highest number of clicks (group 1: black bars, group 2: grey bars and group 3: white bars).

The click distribution was analyzed by counting clicks over a moving window. The window and step length that gave the clearest results (judged by visual inspection of the resulting plots) for the current number of subjects were 1000/500ms.

The click distribution data shows that the subjects highly agree where the synthesis sounds bad. An excerpt of the distribution data is shown in figure 2, where there are five clear peaks of number of clicks. An analysis of what makes group 1 react reveal dense speech passages, often with unexpected input such as foreign names or English terms. Additionaly, group 2 and 3 reacts to bad clips, and sometimes also to odd prosody such as overly long occlusion phases for stops.

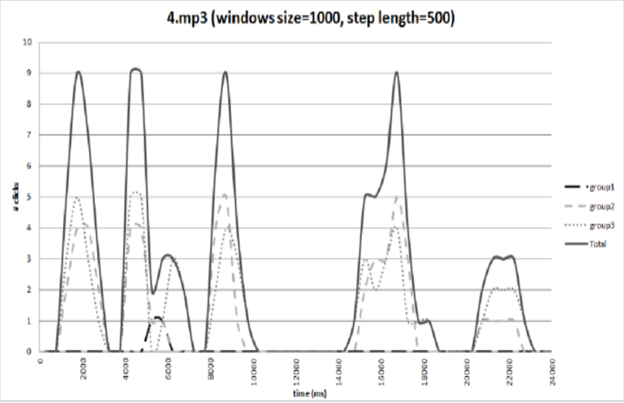


Figure 2. The first 24 seconds of the distribution results of the fourth file.

5. Discussion

The fact that the number of clicks differ so much between group 1 and the other two groups shows that it is possible to ask the subjects for different information; primarily it is important to investigate how often (and in which contexts) the speech synthesis is unintelligible, secondarily to know when the listeners are distracted by any mistake of the synthesis. The similar numbers of clicks by group 2 and group 3 could possibly mean that everything that isn’t entirely correct is annoying, but it might simply be the individuals' interpretation of the task. More subjects are needed to clarify this.

Despite the large variation in click counts within group 2 and 3, the distribution data is homogeneous, which bodes well for the continued use of the ARS inspired test method for speech synthesis.

The total listening time, including the practice file, was around 3.5 minutes. Assuming that it takes five minutes to log in and read the instructions, each subject spent less than ten minutes on the test. With 24 subjects, the total time spent on the test was less than 4 hours, of which 1 hour and 24 minutes was active evaluation time. In comparison with a grading test or a preference test, where the subjects can be estimated to evaluate about ten sentences or sentence pairs in ten minutes, the coverage resulting from the click test is huge. Furthermore, using a web-based test where the results are easily downloaded makes it a rather quick process to put together the results in different constellations.

The subjects' responses to the test method were mainly positive. Some of them thought it was stressful, while others thought it was not. A general concern of the subjects was the reaction time.

6. References

Benoit, C. (1989). Intelligibility Test for the Assessment of French Synthesisers Using Semantically Unpredictable Sentences. In SIOA-1989, Vol.2, 27-30.

Black, A. and Tokuda, K. (2005). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In Proc. Interspeech 2005.

Edlund, J. (2011). In search of the conversational homunculus. Doctoral Thesis. KTH, Stockholm.

Heuven, V.J. van & Bezooijen, R. van (1995). Quality evaluation of synthesized speech. In W.B. Klein, K.K. Paliwal (eds), Speech coding and synthesis (pp. 707–738). Amsterdam: Elsevier Science.

Entity Recognition of Pharmaceutical Drugs in Swedish Clinical Text

Sidrat ul Muntaha[1], Maria Skeppstedt[1], Maria Kvist[1,2], Hercules Dalianis[1]

[1]Dept. of Computer and Systems Sciences (DSV), Stockholm University, Forum 100, 164 40 Kista, Sweden
[2]Dept. of clinical immunology and transfusion medicine, Karolinska University Hospital, 171 76 Stockholm, Sweden
simu9349@dsv.su.se, mariask@dsv.su.se, maria.kvist@karolinska.se, hercules@dsv.su.se

Abstract

An entity recognition system for expressions of pharmaceutical drugs, based on vocabulary lists from FASS, the Medical Subject Headings and SNOMED CT, achieved a precision of 94% and a recall of 74% when evaluated on assessment texts from Swedish emergency unit health records.

1. Introduction

A patient's health and treatment progress is documented in the clinical record in the form of structured data as well as in the form of narrative text. The data documented in narrative form is difficult to use for e.g. structured summarization, advanced search, statistical analysis and data mining. To be able to use narrative information for these purposes, automatic information extraction tools are called for to retrieve relevant information from free text. (Meystre et al., 2008)

An important part of the health record is documentation of a patient's medication. Automatic text summarization of clinical notes, including parts reasoning about medication, would enable clinicians to form a quick overview, also of records with long and detailed patient histories. Documentation of medication in health records could also be used for mining for new knowledge on pharmaceutical drugs used in health care, e.g. knowledge of adverse drug reactions caused by medication.

The first step for extracting information on medication, both for the purpose of summarization and for text mining, is to automatically recognize drugs that are mentioned in the clinical text. The aim of the work presented here is to study automatic recognition of pharmaceutical drugs mentioned in Swedish clinical text.

2. Method

The general approach of this study was to recognize mentions of drugs using a rule-based matching of clinical text to vocabulary lists, and evaluate this matching on annotated text data.[1]

2.1 Annotation

The data used for evaluation was clinical text annotated for mentions of pharmaceutical drugs. Generic substances, e.g. 'Paracetamol', and pharmaceutical drug names, e.g. 'Alvedon', as well as more general terms denoting medication, e.g. 'smärtstillande' ('pain killer') were annotated.

Free text in the assessment part of clinical notes from an emergency unit of internal medicine at Karolinska University Hospital was used. The texts are part of the Stockholm EPR Corpus (Dalianis et al., 2009) which contains electronic patient records written in Swedish. The same texts were previously used in a study focusing on clinical findings and body structures (Skeppstedt et al., 2012). The annotation had been carried out by a senior physician, using the annotation tool Knowtator (Ogren, 2006).

2.2 Vocabulary lists

The vocabulary for pharmaceutical drugs (25,161 unique expressions) was retrieved from three main sources: The Swedish version of MeSH, Medical Subject Headings (Karolinska Institutet, 2012), the Swedish translation of SNOMED CT (IHTSDO, 2008) and FASS, Farmaceutiska specialiteter i Sverige (FASS, 2012), which provides detailed about approved pharmaceutical drugs in Sweden.

From MeSH, terms in the category *pharmacologic-substance* (2,554 terms) as well as in the category *antibiotic* (239 terms) were used. From SNOMED CT, terms under the main category node *pharmacuetical* (16,977 terms) were used. From FASS, a list of Swedish product names for drugs (7,056 terms) as well as a list of classifications (5,062 terms) were used (NPL, 2011). The FASS terms for classifications of drugs, also contains a few very general terms, and to avoid false positives, terms in this list that were also included in the Swedish non-medical corpus Parole (Gellerstam et al., 2000) were therefore removed.

2.3 Matching to lists

Information in health records is often expressed using abbreviations, medical jargon or misspellings. This writing style has the advantage of quick recording, but makes it more difficult to process by a natural language processing system. As a consequence, an exact match to vocabulary lists might not be sufficient. Therefore, apart from exact string matching, the Levenshtein distance algorithm was used for comparing the clinical text to the terms in the vocabulary list.

The Levenshtein distance is a measure of similarity/distance between two strings, defined as the number of deletions, insertions and substitutions that are needed to transform one string to the other. Experiments were carried out in which expressions that had a Levenshtein distance of one or a Levenshtein distance of two from a term in the vocabulary lists were considered as a matching expression.

[1] The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

The automatic matching was evaluated against the annotated text, using the conll 2000 script (CONLL, 2000).

3. Results

The matching methods were evaluated on the annotated data, and precision, recall and F-measure were calculated. The results are shown in Table 1. A total of 580 mentions of drugs were present in the evaluation data, consisting of 26,011 tokens.

Method	Precision (CI)	Recall (CI)	F-score
Exact match	0.51 (± 0.03)	0.72 (± 0.04)	0.60
Excl. parole	0.94 (± 0.02)	0.74 (± 0.04)	0.83
Lev dist. 1	0.91 (± 0.03)	0.74 (± 0.04)	0.82
Lev dist. 2	0.89 (± 0.03)	0.75 (± 0.04)	0.81

Table 1: Precision, recall and F-score of the matching methods: 'Exact string match', 'Exact match, but words occurring both in classification list and Parole removed', 'Levenshtein distance of 1' and 'Levenshtein distance of 2'. For precision and recall a 95% confidence interval is provided.

4. Discussion

Just above 70% of the words and expressions for drugs were found using exact string matching. The Levenshtein distance matching method did not result in an improvement of recall, but only in decreased precision, which indicates that misspellings are not a common source of error when performing string matching of drugs.

4.1 Error analysis

That misspellings were rare, was also confirmed by the error analysis of the unmatched words. Also abbreviations were few among the false negatives.

Instead, compound words accounted for a large number of unmatched drug expressions, e.g. 'furixbehandling' ('furix treatment'), as well as expressions denoting drugs that were expressed with the effect of the drug or the disease for which it is given e.g. 'blodförtunnande' ('blood thinners') and 'hjärtsviktsmedicinering' ('heart failure medication'). Swedish is a language full of compound words, which provides special difficulties in building/porting tools.

Among the false positives were the term 'läkemedel' ('pharmaceutical') and expressions denoting narcotics.

4.2 Related work

When evaluating vocabulary-based entity recognition of drugs on text in discharge letters, a precision of 95% and a recall of 93% was achieved by Kokkinakis and Thurin (2007). That better results were achieved by Kokkinakis and Thurin (2007) might be due to that different rule-based approaches were used, but it may also be due to different types of evaluation data (discharge letters often have a more formal writing style than assessment notes). A part of the difference could perhaps also be explained by that a more wide definition of what expressions denote a pharmaceutical drug was used in the present study, compared to the study by Kokkinakis and Thurin.

5. Conclusion and future work

The vocabulary-based recognition of pharmaceutical drugs evaluated in this study identified more than 70% of the expressions for drugs in the free text of health records. Since compound words were frequent among the false negatives, compound splitting could be applied to improve results. Also additional methods ought to be applied, such as machine-learning-based recognition of drugs, which has been used by e.g. Wang and Patrick (2009). The vocabulary-based method developed for this study could, however, serve as a baseline method, and more importantly, the method evaluated here could also serve as one of the key features for such a machine learning system.

6. References

CONLL. 2000. CoNLL-2000. http://www.cnts.ua.ac.be/conll2000/chunking/, Accessed 2011-10-09.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR*.

FASS. 2012. Sök läkemedel. http://www.fass.se/LIF/produktfakta/sok_lakemedel.jsp. Accessed 2012-08-27.

Martin Gellerstam, Yvonne Cederholm, and Torgny Rasmark. 2000. The bank of Swedish. In *Proceeding of LREC*, pages 329–333.

IHTSDO. 2008. SNOMED Clinical Terms User Guide, July 2008 International Release. http://www.ihtsdo.org. Accessed 2011-01-24.

Karolinska Institutet. 2012. Hur man använder den svenska MeSHen (In Swedish, translated as: How to use the Swedish MeSH). http://mesh.kib.ki.se/swemesh/manual_se.html. Accessed 2012-03-10.

Dimitrios Kokkinakis and Anders Thurin. 2007. Identification of entity references in hospital discharge letters. In *Proceedings of NODALIDA*.

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–144.

NPL. 2011. NPL (Nationellt produktregister för läkemedel) Review and verify product information. https://npl.mpa.se/mpa.npl.services/home2.aspx. Accessed 2011-10-28.

Philip V. Ogren. 2006. Knowtator: A Protégé plug-in for annotated corpus construction. In *Proceedings of HLT-NAACL*.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based entity recognition and coverage of snomed ct in swedish clinical text. In *Proceedings of LREC*.

Yefeng Wang and Jon Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 42–49.

Classifying Pluricentric Languages: Extending the Monolingual Model

Marcos Zampieri[1], Binyam Gebrekidan Gebre[2], Sascha Diwersy[3]

University of Cologne[1,3], Max Planck Institute for Psycholinguistics[1]
Albertus-Magnus-Platz 1, 50931, Cologne, Germany[1,3]
Wundtlaan 1, 6525 XD, Nijmegen, Holland[2]
mzampier@uni-koeln.de, bingeb@mpi.nl, sascha.diwersy@uni-koeln.de

Abstract

This study presents a new language identification model for pluricentric languages that uses n-gram language models at the character and word level. The model is evaluated in two steps. The first step consists of the identification of two varieties of Spanish (Argentina and Spain) and two varieties of French (Quebec and France) evaluated independently in binary classification schemes. The second step integrates these language models in a six-class classification with two Portuguese varieties.

1. Introduction

Automatic Language Identification is a well-known research topic in NLP. State-of-the-art methods consist of the application of n-gram language models to distinguish a set of languages automatically. One of the first studies to apply n-gram models to language identification was Dunning (1994) and more recent experiments include Martins and Silva (2005) and Lui and Baldwin (2012).

Martins and Silva use internet data to identify a set of 12 languages and they report results ranging from 80% to 99% accuracy depending on the language. Lui and Baldwin report 91.3% for a set of 67 languages using Wikipedia data. Their software, called *langid.py*, claims to have language models for 97 languages by using various data sources.

These two examples do not take language varieties into account. Pluricentric languages are always modeled as a unique class and this is one reason for the good results these n-gram methods report. The level of success is usually high when classification includes languages which are typologically not closely related (e.g. Finnish and Spanish) and languages with unique character sets (e.g. Hebrew).

1.1 Classifying Varieties

Only recently the automatic identification of language varieties has received more attention. A few studies have been published such as Ljubesic et al. (2007) for the former Serbo-Croatian varieties, Huang and Lee (2008) for Mainland and Taiwanese Chinese and Trieschnigg et al. (2012) for Dutch dialects.

These methods aim to distinguish varieties and to our knowledge none of them has yet been integrated into a broader language identification setting. Here we try to replicate the experiments carried out by Zampieri and Gebre (2012) for Brazilian and European Portuguese on two varieties of French and two of Spanish. Subsequently, we integrate these four new language models into a six-class classification scheme.

2. Methods

We compiled four journalistic corpora containing texts from each of the four varieties. To create comparable samples, we retrieved texts published in the same year (2001 for French and 2008 for Spanish) and all metainformation and tags were removed. The length of texts in the corpora varies and as language identification benefits from longer texts, we only used texts consisting of up to 300 tokens.

Location	Newspaper	Year
Argentina	La Nación	2008
Spain	El Mundo	2008
France	Le Monde	2001
Quebec	Le Devoir	2001

Table 1: Corpora

The identification method works on three different aspects of language: orthography, lexicon and lexico/syntax. For the orthographical differences, we used character n-grams ranging from 2 to 6-grams. At the lexical level we used word uni-grams and finally, to explore lexico-syntactical differences, word bi-grams were used. The language models were calculated with Laplace probability distribution:

$$P_{lap}(w_1...w_n) = \frac{C(w_1...w_n) + 1}{N + B} \qquad (1)$$

In equation number 1: C is the count of the frequency of w_1 to w_2 in the training data, N is the total number of n-grams and B is the number of distinct n-grams in the training data. For probability estimation, we used the log-likelihood function:

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \qquad (2)$$

N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability for each of the language models. The language model with higher probability determines the identified language of the text.

3. Results

Evaluation was done using each of the feature groups in a set of 1,000 documents sampled randomly. The sample contains 50% of the texts from each variety and it is divided into 500 documents for training and 500 for testing.

3.1 Binary Classification

We report results in terms of accuracy for binary classification as seen in table 2.

Feature	AR x ES	FR x QU
Word 1-grams	0.948	0.968
Word 2-grams	0.894	0.956
Character 2-grams	0.898	0.956
Character 3-grams	0.948	0.990
Character 4-grams	0.944	0.968
Character 5-grams	0.962	0.960
Character 6-grams	0.960	0.934

Table 2: Binary Classification

The results suggest that, on average, the French corpora have a stronger variation than the two varieties of Spanish. French scores were higher in most groups of features except character 5 and 6-grams. The results for French and Spanish are, however, lower than those obtained by Portuguese, which reached 0.998 accuracy for character 4-grams.

3.2 Identifying Varieties or Identifying Newspapers?

Studies on identification of language varieties use standard corpora sampled from newspapers and magazines (Huang, 2008). They do not, however, address the question of textual genres and stylistics that underlie these samples. Newspapers and magazines could contain distinctive features that influence the performance of the classifiers. To explore this variable we carried out a controlled experiment using two newspapers from Spain. A corpus with texts from *El País*, published in 2008, was compiled and classified with the *El Mundo* corpus.

Feature	Mundo x País	Difference
W 1-grams	0.614	-33.4%
W 2-grams	0.498	-39.6%
C 2-grams	0.658	-24.0%
C 3-grams	0.654	-29.4%
C 4-grams	0.728	-21.6%
C 5-grams	0.688	-27.4%
C 6-grams	0.564	-39.6%

Table 3: El Mundo x El Pais

Results are 21.6% to 39.6% worse than the classification of Argentinian and Peninsular Spanish. In one case, word bi-grams, the classification result is lower than the 50% baseline expected for binary classification. The poor results obtained suggest that the language models applied here are actually distinguishing the varieties and that the text types and genres do not substantialy influence the algorithm's choice.

3.3 Multilingual Classification

To evaluate the classification model we integrated the four language models described so far: Spain, Argentina, France and Quebec, with two Portuguese varieties: Brazil and Portugal (Zampieri and Gebre, 2012). The results for this six-class classification model are presented in terms of Accuracy, Recall, Precision and F-Measure.

Feature	A	R	P	F
W 1-grams	0.917	0.917	0.905	0.911
W 2-grams	0.878	0.878	0.866	0.872
C 2-grams	0.898	0.898	0.880	0.889
C 3-grams	0.947	0.947	0.933	0.940
C 4-grams	0.910	0.910	0.890	0.899
C 5-grams	0.924	0.924	0.905	0.915
C 6-grams	0.935	0.935	0.932	0.933

Table 4: 6-Class Classification

4. Conclusion and Future Work

Studies on language indentification neglect pluricentric languages. We therefore presented a language identification model focusing on language varieties. The best results on the binary classification were obtained by using character n-grams. For Argentinian and Peninsular Spanish, 0.962 using 5-grams and for Quebec and Hexagonal French, 0.990 using 3-grams. The six-class model reached 0.947 accuracy and 0.940 f-measure using character 3-grams.

This work shed light on two areas. First, it shows that language identification models may include language varieties without substantial loss of performance. This should help to increase performance in NLP applications such as spell checking and MT systems. The second area is contrastive linguistics. Pluricentric languages are often the object of study due to their variation in grammar and syntax. Classification experiments such as this can provide a quantitative overview on how varieties diverge and converge.

Further experiments include the integration of these models into broader classification schemes (up to 20-fold) and the use of more knowledge-rich features such as POS bi-grams, to measure the extent to which these varieties differ in terms of grammar.

5. References

T. Dunning. 1994. *Statistical Identification of Language* Technical Report MCCS-94-273 New Mexico State University

C. Huang; L. Lee. 2008 Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity *Proceedings of PACLIC 2008* p. 404-410

N. Ljubesic; N. Mikelic; D. Boras. 2007. Language Identificaiton: How to Distinguish Similar Languages? *Proceedings of the 29th International Conference on Information Technology Interfaces*

M. Lui; T. Baldwin. 2012 langid.py: An Off-the-shelf Language Identification Tool *Proceedings of the 50th Meeting of the ACL* p. 25-30

B. Martins; M. Silva. 2005. Language Identification in Web Pages *Proceedings of the 20th ACM Symposium on Applied Computing (SAC)* Santa Fe, EUA. p. 763-768

D. Trieschnigg; D. Hiemstra; M. Theune; F. de Jong; T. Meder. 2012. An exploration of language identification techniques for the Dutch Folktale Database. *Proceedings of LREC2012*

M. Zampieri; B. G. Gebre. 2012. Automatic Identification of Language Varieties: The Case of Portuguese *Proceedings of KONVENS2012*. Vienna, Austria. p. 233-237

Vocabulary and Syntax in Gender Classification

Niklas Zechner

Department of Computing Science,
Umeå University
zechner@cs.umu.se

1. Introduction

There are many different kinds of text classification. One can identify specific authors, subjects, styles, or attributes of the authors, such as age, native language, age, education level, and so on. In this article, I will focus on the elusive differences of gender. One obvious method in most aspects of text classification is to look at word frequencies. However, this approach is not without problems; word frequencies are strongly affected by such things as subject, and can not be expected to be the same for a given author writing different texts. For this reason, one interesting approach is to look at syntax instead. Syntax is strongly affected by style, context, and level of formality, but, one would suspect, less affected by subject (Stamatatos, 2009). In this study we look at novels, all written in a timespan of six years, so it seems reasonable to believe that the context and level of formality should be about the same.

2. Method

The data used comes from two corpora from Språkbanken, consisting of Swedish novels: ‘Bonniersromaner I’, consisting of 69 novels written between 1976 and 1977, and ‘Bonniersromaner II’, consisting of 60 novels written between 1980 and 1981. Together, they include 10.9 million words written by 108 authors. The corpora are pre-parsed and annotated.

The novels are divided into blocks of a number of words, and each block is treated as a separate text. Each block is marked with the gender of the author, and those which are written by more than one person are removed. I have looked at three types of features. First, frequencies of common tokens, i.e. words as they are written, treating inflected forms as different. Second, frequencies of common lemmas, i.e. words changed to their dictionary form. Third, frequencies of common parent-child subtrees in the dependency tree; the feature identifies the parts of speech of the two words in such a subtree.

After a series of preliminary experiments, it appears that the best accuracy is achieved using large blocks, consisting of 10 000 words. Using even larger blocks decreases the accuracy, as the blocks get close to the lengths of some novels. For the best effects, I use a large number of features, setting the limit at 1000 of each of the features. It should be noted however that there are only 596 trees in total.

I use a naive Bayes classifier, which is trained on the majority of the data and then evaluated on the remainder; so far this has proven to be the most effective classifier. There are 1086 blocks in the data. I test the accuracy by two methods; first, using cross-validation with ten folds, and second, using specific test sets. For cross-validation, I also use an arbitrary division into two groups, a sort of ‘false genders’, to see how the accuracies compare. For the specific test sets, I pick sets of four arbitrary authors of each gender, so that the male and female authors are represented by nearly equally large amounts of data. The advantage of this method is that there are no texts in the test set by the same authors as in the training set, so we are more reliably identifying gender and not author.

For each of the features, I analyse the relative difference between men and women, to see which features are the most prominent. To avoid including very rare words - which would very likely come from a single book - I define prominence for this purpose as the difference multiplied by the average frequency of the feature. Thus the top items on the list are those features which are both common and vary greatly between men and women.

3. Results

It is clear that lemmas and tokens are more effective than trees, to the point where using only lemmas or only tokens is about as effective as using all three. Lemmas and tokens give very similar results, with lemmas being possibly a little bit more effective, so I focus on testing either only lemmas or only trees.

For lemmas, cross-validation gives 85% accuracy, and the arbitrary division gives 76%. With specific test sets, the average is 73%, with a standard deviation of 12%.

For trees, cross-validation gives 76% accuracy, and the arbitrary division gives 67%. With specific test sets, the average is 63%, with a standard deviation of 17%.

Table 1 shows the most prominent differences for lemmas. The first two columns show the actual lemma, and the English translation. The third column shows the difference, in terms of how many percent more often men use the lemma than average; thus, a negative value means that women use the lemma that many percent more than the average. The last column shows the average frequency of the lemma, per block (of 10 000 words).

Table 2 shows the most prominent differences for trees, with the codes used by Språkbanken (2012). We see for example that women are more likely to use pronouns as arguments of verbs (VB-PN) whereas men are more likely to use nouns as arguments of verbs (VB-NN).

lemma	transl	diff	freq
hon	she	-35.47	146.99
en	a, one	7.27	320.78
jag	I	-8.22	231.98
du	you (sg.)	-19.32	73.85
inte	not	-9.50	114.92
vi	we	19.40	55.09
i	in	5.89	174.10
och	and	3.11	309.67

Table 1: The most prominent lemmas.

tree	diff	freq
VB-PN	-6.77	931.61
PP-NN	4.74	532.84
VB-NN	4.85	505.49
NN-DT	7.28	331.52
KN-NN	11.47	193.75
NN-JJ	8.28	258.13

Table 2: The most prominent trees.

4. Analysis

For the lemmas, it seems clear that the prediction is significantly better than chance. For the trees, the results are more ambiguous. More tests would help determine if the difference is statistically significant.

If we look at the separate features, a statistical analysis shows that many of them are very unlikely to be chance differences. But that is under the assumption that the authors – and books – are good representations of their genders.

One problem that needs to be taken into account is that the program might identify authors or books, and then implicitly use that to identify gender. For example, if there is a male author writing about bears, we would see a strong correlation between the word 'bear' and being male. This seems like the most likely reason why the cross-validated results are higher, even for the arbitrary division. For the specific test sets, this should not be a problem.

There are many stereotypes about genders and how they write, and it is easy to draw conclusions from the data which fit them. But it is not easy to know which of these differences are genuine, and which are coincidences.

The very top item of the list is the word for 'she', which is hardly surprising. It seems very reasonable that women are more likely to write about other women. Lower on the list, we can also see that women are more likely to use the words for 'mother' and 'child'.

Several other pronouns also end up near the top of the lists, and generally they are used more by women. This is also reflected in the second table, where pronouns as arguments are more common among women, and nouns as arguments more common among men. The major exception is 'we', which is used more commonly by men.

Koppel et al. (2002) finds similar results for English texts. They find that among the most prominent male words are articles like 'a' and 'the', and among the most prominent female words are 'she' and other pronouns.

Chung and Pennebaker (2007) looks at pronoun use in Japanese and US English texts, and finds that Japanese texts are more likely to use 'I', and the American texts are more likely to use 'we'. They find this surprising, since the assumption is that the US writers are more individualistic, but they speculate that 'we' signals equality, whereas the self-deprecation typical of a hierarchical society leads to expressions like 'they and I'. We see that women are more likely to use 'I', and men more likely to use 'we'; this could similarly be seen as women being more self-deprecating.

From the tree frequencies, we find that men are more likely to write longer sentences. The list of lemmas also shows that men more frequently use words like 'and', 'or', 'this', 'also'. This might suggest that they use more complex syntax, more dependent clauses, etc.

One stereotype is that women would be more likely to talk about relations, emotions and communication. We can make a list of words which are related to that concept, and see what the statistics say. Judging what is an emotion word is of course subjective, but I find the following words among those with a prominence higher than 50: 'say', 'want', 'know', 'think', 'have an opinion', 'believe', 'feel', 'face', 'love' (as verb), 'home', 'phone' (as verb), 'understand', 'speak', 'laugh', 'talk', 'smile', 'nice, kind', 'mean' (as verb), 'happy', 'write', 'each other', 'love' (as noun), 'help', 'cry', 'together'.

Out of these 25 words, 1 is used more by men, 'write' at 11%; 1 is used between 0 and 10% more by women, 'speak' at 9%; 14 are used between 10 and 20% more by women; 7 are used between 20 and 30% more by women; 2 are used more than 30% more by women; 'nice, kind' at 37% and 'love' (as verb) at 40%.

Note that these percentages are differences from the *average*; thus, women in the sample used the word 'love' 40% more than average, i.e. 135% more than men. Furthermore, 'yes' and 'no' are also more commonly used by women, which points to more use of direct speech. If this is the case, it could also explain the shorter sentences.

It is worth noting that these words probably correlate strongly with each other, so with a small number of authors there is the possibility that all the female authors coincidentally wrote about these particular things. However, it is difficult to argue that 108 is a small number of authors in this context. Future work could aim at validating these patterns by setting up and testing formal hypotheses.

5. References

C. Chung and J. Pennebaker. 2007. The psychological functions of function words. In K. Fiedler, editor, *Social Communication*. Psychology Press.

M. Koppel, S. Argamon, and A. R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17.

Språkbanken. 2012. MSD-taggmängd. Accessed at 12-08-01. http://spraakbanken.gu.se/korp/msdtags.html.

E. Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60.

Stagger: A modern POS tagger for Swedish

Robert Östling

Department of Linguistics
Stockholm University
SE-106 91 Stockholm
robert@ling.su.se

Abstract

The field of Part of Speech (POS) tagging has made slow but steady progress during the last decade, though many of the new methods developed have not previously been applied to Swedish. I present a new system, based on the Averaged Perceptron algorithm and semi-supervised learning, that is more accurate than previous Swedish POS taggers. Furthermore, a new version of the Stockholm-Umeå Corpus is presented, whose more consistent annotation leads to significantly lower error rates for the POS tagger. Finally, a new, freely available annotated corpus of Swedish blog posts is presented and used to evaluate the tagger's accuracy on this increasingly important genre. Details of the evaluation are presented throughout, to ensure easy comparison with future results.

1. Introduction

The task of syntactic disambiguation of natural language, frequently referred to as part of speech (POS) tagging, aims to annotate each word token in a text with its part of speech and (often) its morphological features.

I have implemented a new, freely available POS tagging system for Swedish, named *Stagger*,[1] and used it to evaluate recently developed tagging algorithms on Swedish, as well as the effects of improved corpus annotation on POS tagging accuracy.

2. Data

Two corpora were used to evaluate the accuracy of the POS tagger: an updated version of the Stockholm-Umeå Corpus, and a new corpus of Swedish blog texts.

2.1 Stockholm-Umeå Corpus

The Stockholm-Umeå Corpus (SUC) is a balanced and POS-annotated corpus of about one million words of Swedish text, which was originally developed at the universities of Stockholm and Umeå during the 1990s. Its most recent release (Gustafson-Capková and Hartmann, 2008) has become a de-facto standard for Swedish POS tagging research.

Due to the size of the corpus, multiple annotators have been used, and annotation (in)consistency is an issue. Källgren (1996) explored tagging errors in an earlier version of the corpus, and found that 1.2% of the words sampled contained POS annotation errors. Forsbom and Wilhelmsson (2010) corrected over 1500 errors in SUC 2.0, mostly in common, polysemous grammatical words, and found that this results in a small but significant improvement in POS tagger accuracy.

We have included the changes of Forsbom and Wilhelmsson (2010), as well as over 2500 other changes to the annotation, into version 3.0 of SUC.[2]

2.2 Swedish blog texts

The language in so-called *user-generated content*, written by non-professionals in for instance blog posts or online forum posts, may differ considerably from traditional written language and poses a challenge to many Natural Language Processing applications, including POS tagging (Giesbrecht and Evert, 2009).

In order to evaluate the current POS tagger on user-generated content in Swedish, a small corpus (8 174 tokens) of blog texts was compiled and manually annotated with SUC-compatible POS tags and named entities. The corpus is freely available for download from the Stockholm University website.[3]

2.3 Unannotated data

For semi-supervised training, Collobert and Weston (2008) embeddings were induced from a corpus of about two billion tokens of Swedish blog texts.

2.4 Lexicon

In addition to the vocabulary in the training data, the SALDO lexicon of Swedish morphology (Borin and Forsberg, 2009) is used as a POS tag lexicon. For known words, only POS tags occurring with the word in either the training data or the SALDO lexicon are considered. For unknown words, all POS tags that occur with a token of the same type (e.g. number, emoticon or letter sequence) are considered.

3. Method

The tagger uses a feature-rich model, based on the averaged perceptron tagger of Collins (2002).

A basic feature set similar to Collins' is used. Details are omitted due to space limitations, but are documented in the software package. In addition, 48-dimensional Collobert and Weston (2008) embeddings (C&W) were used as features in one tagger configuration. Each word can then be

[1] http://www.ling.su.se/stagger

[2] More information and instructions for obtaining the corpus can be found at: http://www.ling.su.se/suc

[3] http://www.ling.su.se/sic

Table 1: POS tagging accuracy in percent, with figures in bold significantly better than the others in the same column.

Configuration	**SUC2**	**SUC3**	**Test3**	**Blogs**
–	95.86	96.04	96.58	91.72
SALDO	96.32	96.52	**96.94**	**92.45**
SALDO+C&W	**96.40**	**96.57**	**96.94**	92.10

represented by a 48-dimensional vector, reflecting distributional (and indirectly syntactic and semantic) properties of the word.

4. Results

Using SUC, two evaluations were performed: 10-fold cross validation, and another using the training/development/test split in SUC 3.0. In the cross-validation, the 500 files were sorted alphanumerically and numbered from 0 to 499. Fold i (0..9) uses files k where $k \equiv i \pmod{10}$ for testing, and the remaining 450 files for training. A held-out set of 10 files in the training set are used to determine the number of training iterations.

In the non-cross validations, the development set (2%) of SUC 3.0 was used for this purpose, and the test set (2%) used for estimating the final accuracy.

Table 1 shows the results of the evaluation. **SUC2** and **SUC3** are cross-validations using SUC 2.0 and 3.0, respectively. **Test3** uses the training/development/test sets of SUC 3.0, and **Blogs** uses the training and development sets of SUC 3.0 for training, and the annotated blog corpus for testing.

Forsbom and Wilhelmsson (2010) found that a subset of the changes to SUC explored in this work led to a significant reduction in errors made by a POS tagger, and as expected this effect was larger after more errors were corrected.

The evaluation also demonstrated the importance of using a good lexicon, where the SALDO lexicon of Swedish morphology made a great contribution to tagging accuracy.

Finally, Collobert & Weston embeddings were shown to improve tagging accuracy by a small but significant[4] amount in the cross-evaluation, similar to what Turian et al. (2010) showed for other NLP tasks. Surprisingly, given the fact that the embeddings were computed from an unannotated blog corpus, the accuracy on the annotated blog corpus is instead significantly *lower* with C&W embeddings. However, since there are only three authors represented in the blog corpus, it would be risky to draw too general conclusions on the basis of this result.

5. Related work

Sjöbergh (2003) evaluated seven different POS tagging systems for Swedish through ten-fold cross-validation on SUC 2.0, where accuracies ranged between 93.8% and 96.0% for single systems (Carlberger and Kann, 1999, being the best), and a voting combination of all taggers reached 96.7%. However, since the details of his evaluation were not published, and he used a larger training data set in each fold (95%) than the present study (90%), our respective accuracy figures are not directly comparable.

6. Acknowledgments

Thanks to those who found and corrected the thousands of annotation errors in SUC 2.0 that have been fixed in SUC 3.0: Britt Hartmann, Kenneth Wilhelmson, Eva Forsbom and the Swedish Treebank project. Further thanks to the two anonymous reviewers, who provided useful comments.

7. References

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software–Practice and Experience*, 29:815–832.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP 2002*, pages 1–8.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Eva Forsbom and Kenneth Wilhelmsson. 2010. Revision of part-of-speech tagging in stockholm umeå corpus 2.0. In *Proceedings of SLTC 2010*.

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*.

Sofia Gustafson-Capková and Britt Hartmann, 2008. *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University.

Gunnel Källgren. 1996. Linguistic indeterminacy as a source of errors in tagging. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 676–680, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonas Sjöbergh. 2003. Combining pos-taggers for improved accuracy on swedish text. In *Proceedings of NODALIDA*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

[4] McNemar's test with $p < 0.05$ is used throughout to test for statistical significance.

Author Index