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.