Evaluation of statistical categorization methods for creating specialized vocabulary lists to be used as learning aid

Christian Lindgren
Lund University
Lund, Sweden
ada07cli
@student.lu.se

David Larsson
Lund University
Lund, Sweden
lak03dbo
@student.lu.se

Lars Gustafsson
Lund University
Lund, Sweden
ada10lgu
@student.lu.se

Abstract

This paper examines the possibility of creating a shortcut in category targeted learning of a new language through filtering category word lists using two gold-standard statistical methods: Student's T-test and Chi-squared method. The word lists are compared to each other using only the most frequent words in a large training corpus with coverage of the test corpus as the main measurement. The results are rather disappointing and the coverage of the filtered lists don't differ significantly from using only a list sorted by frequency. Studies argue however that the statistical methods used produce a rather large amount of false positives and future work should therefore examine other methods as presented by linguistic studies.

1 Introduction

The purpose of this paper is to examine whether there exists a shortcut to learning a new language with the premise that one is not interested in learning about more than one or a few disciplines (e.g physics, sports, plants etc.) as well as learning enough words to read a general article in that language. The idea is that using categories of words one can filter out a list of significant words to be learned in order to be able to read articles and converse with people about that category. On top of that one must learn a set of general words in order to understand texts and dialogues in that language. Apart from the required grammar, this paper postulates that a higher coverage of the words in a given category corpus leads to a better understanding of a language. This paper's thesis is that a combination of the general words and the category specific words are sufficient to be able to understand a language as long as one does not wander outside the category chosen. The number of words needed to be learned using this method would be lower than the number of words that one might need to learn using a conventional method of learning a language.

1.1 Previous work

Frequency analysis for comparing frequencies of words in corpora has been thoroughly examined before [1] [2] [3] [4]. Rayson and Garside presents a variation of the Chi-squared method for filtering out significant words and sorting them according to significance. Their results are promising and the main reason for choosing their method as one of the filters in this study.

2 Methodology

The method used is frequency analysis on training corpora to predict the most frequent words in a test corpus. Given two training corpora, one that represents the language in general and one that is category specific, the goal is to cover an as large proportion of the text in the test corpus as possible.

The words are tagged with part-of-speech (POS) tags to prevent ambiguity. For instance a corpus on the category *golf* may contain a high frequency of the *noun* green, while a corpus on the category *colours* may contain a high frequency of the *adjective* green. In order of being able of separate those words POS tagging is needed.

The task is narrowed down to regard category specific prediction, meaning that a corpus for a specific category is required to make predictions on that very category.

The first step is to get a large corpus which contains as general content as possible, which will be called corpus A in this paper. Furthermore one needs a category-specific corpus, here called corpus B.

The second step is to POS tag the corpora and then simply sort the words by their frequencies.

Four different methods for frequency analysis are examined.

1. Simply using only words from corpus A, sorted by frequency as a filter, starting with the most frequent.

2. Simple category specific method, using the same method as in 1, but has a window between word X_1 and X_2 where it uses corpus B instead.

3. Students T-test, running a T-test between corpus A and B to get significant words with 95% accuracy into a new corpus B_1 and then continue as in method 2 replacing B with B_1.

4. Chi-squared, the same as method 3 but replacing 95% accurate T-test with Chi-squared statistics, sorted by significance according to the Chi-squared value.

All parts of speech except nouns, verbs, adjectives and adverbs are filtered out in order to examine only words containing actual information. It's the belief of the authors that words like "and" as well as "on"[1] don't hold any significance at all for a category specific vocabulary. The initial usage of the corpus A is to get rid of common words like, "is", "contains", "exists" etc. since these words indeed are very common in the test corpus but are of less relevance for the category specific list of words.

In this paper the content of the Swedish Wikipedia as of 2013-11-01 is used as corpus A and articles linked under a specific category on Wikipedia as corpus B. The test corpus consists of several news articles connected to the same category as corpus B.

The system is constructed by four separate subsystems: Data to categories, tagging, evaluating and presenting. The base files used by the system are an XML-dump from Wikipedia containing all articles in a given language and the "Data to categories" produces one file for each requested category and one for the entire text now stripped down to raw text without any formatting. After that the files are used by the tagging software and it calculates the frequency of each word in every file and returns a list with all words and the number of occurrences for each word. Then the program evaluates the lists and compares them to produce the final lists filtered according to the methods explained above. Lastly the data is presented both as a list and a wordcloud[2].

The result is presented as a percentage of the maximum possible coverage for a given amount of words (X_N), a window (X_1 to X_2) and a category, as well as the over all coverage for the same parameters.

2.1 Method 1

Given a list of all the words in corpus A, sorted by frequency, pick the first in the list and iterate X_N times.

For graphs see Appendix A.

Method 1		
Category	Coverage	
	Over-all	Relative optimal
Kampsport	52.6 %	77.0 %
Algoritmer	58.8 %	79.5 %

Table 1: Values for method 1 given $X_N = 700$.

2.2 Method 2

Given the same list as in Method 1, perform the same iteration but for X_1 times, then switch over to corpus B and continue for $X_2 - X_1$ times, finally switching back to corpus A and continue $X_N - X_2$ times (always skipping already picked words).

For graphs see Appendix B.

Method 2		
Category	Coverage	
	Over-all	Relative optimal
Kampsport	59.8 %	87.7 %
Algoritmer	62.3 %	84.3 %

Table 2: Values for method 2 given $X_1 = 200$, $X_2 = 500$ and $X_N = 700$.

2.3 Method 3

Given the corpora A and B_1, where B_1 is corpus B filtered by T-test according to a method described in [5], where the standard deviation is approximated by the frequency itself, and then sorted by

[1] Even though the study is done on the Swedish language we use English words for the readers convenience.

[2] Wordcloud - a graphical representation of a text where the font size of a given word corresponds to the frequency of the given word in the text.

frequency. The formula

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{1}$$

where

$$s = p(p-1) \tag{2}$$

is used for calculating the t-value as proposed. After that the procedure is identical to method 2.
For graphs see Appendix C.

Method 3		
Category	Coverage	
	Over-all	Relative optimal
Kampsport	59.8 %	87.8 %
Algoritmer	62.3 %	84.3 %

Table 3: Values for method 3 given $X_1 = 200$, $X_2 = 500$ and $X_N = 700$.

2.4 Method 4

Given the corpora A and B a word frequency list B_2 is created as proposed by Rayson and Garside [4]. The method is based upon log-likelihood and chi-squared statistics to create the list using the formulas

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{3}$$

and

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right) \tag{4}$$

Then the same concept as in method 2 of a window is used.

For graphs see Appendix D.

Method 4		
Category	Coverage	
	Over-all	Relative optimal
Kampsport	60.0 %	88.6 %
Algoritmer	61.5 %	83.1 %

Table 4: Values for method 4 given $X_1 = 200$, $X_2 = 500$ and $X_N = 700$.

3 Possible applications

The software is mainly constructed as a tool for learning languages, described in the Introduction, but during the course of the study a few other applications were considered.

3.1 Text categorization

Some experiments were done, not related to the original purpose of the study, to see if one could automatically categorize a given text in to one predetermined category. By use of the Students T-test or Chi-squared the new article or text was given a percentage of similarity to the different texts in the categories.

3.2 Profiling texts

The program can be used to profile a text or texts to present a Wordcloud so its easy to get a general idea what the text is about.

4 External software

Almost all software was developed specifically for this study by the authors but three external tools were also used.

Due to the text extracted from Wikipedia uses a specific markup language a parser was constructed to extract the raw text. The streaming parser, made by the authors, was constructed with focus on speed but lacked in accuracy. Therefore another tool was used. The Wiki-markup filter was made by Peter Exner, a Ph.D. student at the Department of Computer Science, Lund University. By using the new parser an almost 100% success-rate was accomplished.

To achieve some separation of homographs, even though words with the same part of speech will be seen as the same, from the raw text a part-of-speech-tagger was used, called *Stagger*. Stagger [6] is made by the University of Stockholm and based on Collins (2002) averaged perceptron and is one of the best Swedish POS-taggers when it comes to accuracy with about 96.6 percent.

Lastly JDOM was used as an XML parser.

5 Discussion

A few different ”problems” were discovered when manually checking the results. If the text contained the same base word but with different inflections these words were counted as totally separate words. One way to solve this is to reduce all words to their base form before calculating the frequencies. This could dramatically change the significance of some words.

Just as inflections may create several instances of the same word there are some languages that have homographs that are the same part of speech.

This might give a falsely high frequency for some words.

Another thing that was never really discussed when forming the main thesis was that some words may be useful to know for their characteristics even though the words themselves might not be useful. Since the study only evaluates the list of words based on the frequency and no knowledge of the structure of the language this is totally overlooked.

Furthermore grammar is something that this method takes no notice of; the authors sees this tool as an aid when learning a new language. They do understand that learning a new language is as much grammar as learning the words. The system only provides a list of words related to the language. What the pupil has to do is to use a dictionary to figure out the meaning and pronunciation of the given words. One could also argue that it is possible to figure out what meaning of a word is relevant. For example the English word ”bow” have several meanings. When learning the category of Archery its fairly easy to understand that the correct translation to Swedish is ”båge” instead of ”rosett” that is a tied ribbon.

Lastly, T-test and Chi-squared filters might not be the preferred methods when choosing what words are significant [1][2]. Kilgarriff argues that since language is never random, the standard null hypothesis methods are less useful since they produce too many false positives. This might mirror the result of this study where some very common words made the lists of significant words in the different categories. Moreover, Jefrey Lijffijt et al [3] presents two alternatives to the classical statistical methods, *inter-arrival times* and *bootstrapping* which they prove to result in far less false positives than the gold standard methods. Future work for this study might be to test the program with these methods instead to produce even better results with less common words.

6 Conclusions

The lists filtered out by the methods used are intuitively good but lacks an objective measurement of the rate of actual significance the words hold with respect to the structure of language as discussed above. One can however conclude that however good or bad, T-test and chi-squared methods produce largely the same result as simply using the category training data directly for both categories tested.

Furthermore this paper assumes that coverage is more important than the individual words, regarding what words are needed to understand the text. However this might not be the case at all, complementary research is needed in order to conclude whether a human understands a text better with lower coverage and larger amount of key words or if there is a balancing point in-between.

References

[1] Adam Kilgarriff. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276, 2005.

[2] Stefan Th Gries. Null-hypothesis significance testing of word frequencies: a follow-up on kilgarriff. *Corpus linguistics and linguistic theory*, 1(2):277–294, 2005.

[3] Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *Machine Learning and Knowledge Discovery in Databases*, pages 341–357. Springer, 2011.

[4] Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora*, pages 1–6. Association for Computational Linguistics, 2000.

[5] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[6] Robert Östling. Stagger: an open-source part of speech tagger for swedish. *Northern European Journal of Language Technology*, 3:1–18, 2013.

Appendix A

Graphs for method 1.

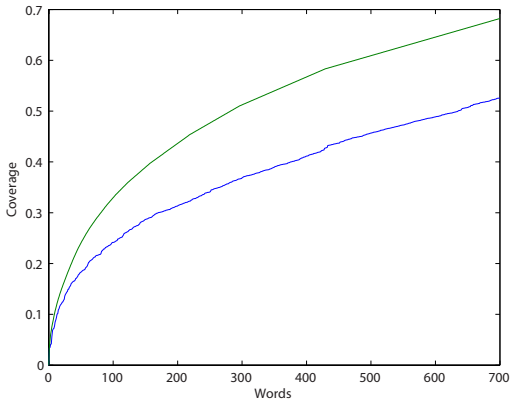

Figure 1: Blue is category ”Kampsport” with Method 1, Green is Optimal from test data.

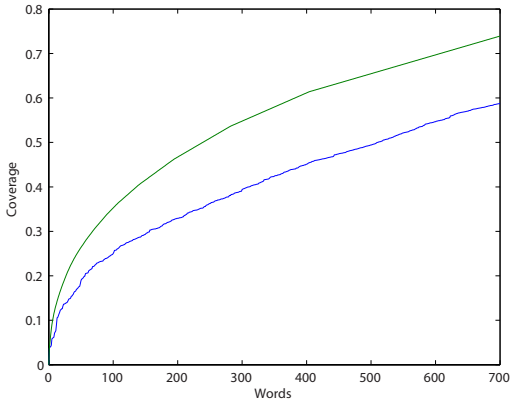

Figure 2: Blue is category ”Algoritmer” with Method 1, Green is Optimal from test data.

Appendix B

Graphs for method 2.

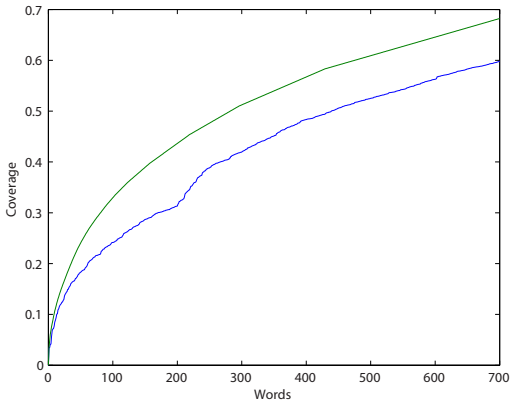

Figure 3: Blue is category ”Kampsport” with Method 2, Green is Optimal from test data.

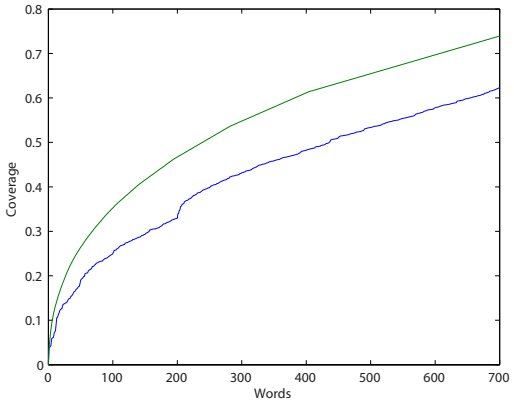

Figure 4: Blue is category ”Algoritmer” with Method 2, Green is Optimal from test data.

Appendix C

Graphs for method 3.

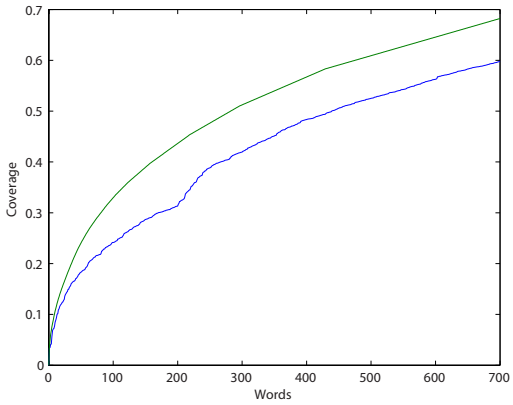

Figure 5: Blue is category ”Kampsport” with Method 3, Green is Optimal from test data.

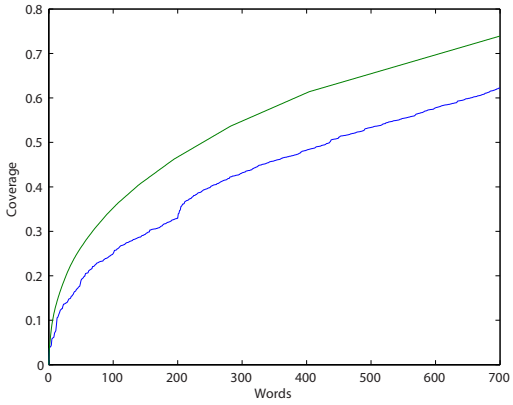

Figure 6: Blue is category ”Algoritmer” with Method 3, Green is Optimal from test data.

Appendix D

Graphs for method 4.

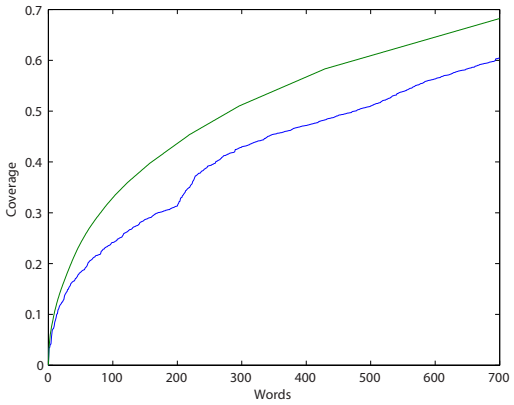

Figure 7: Blue is category ”Kampsport” with Method 4, Green is Optimal from test data.

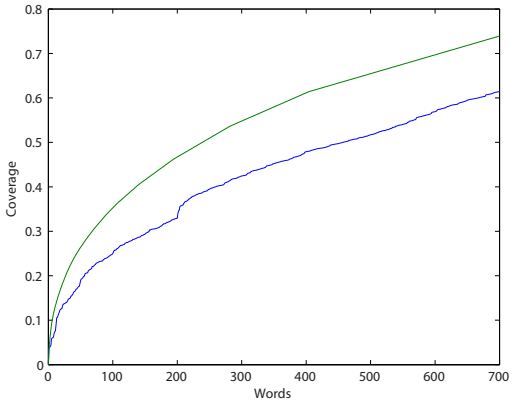

Figure 8: Blue is category ”Algoritmer” with Method 4, Green is Optimal from test data.