Original scientific paper

First Steps Toward Developing a System for

Petra Bago
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
pbago@ffzg.hr

Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: dboras@ffzg.hr

Nikola Ljubešić
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: nljubesi@ffzg.hr

Summary

The aim of this paper is to describe first steps in developing a system for terminology extraction. First a data sample is built from synopses of doctoral theses at the Faculty of Humanities and Social Sciences, University of Zagreb, accepted in the period from 2004 to 2009 written mostly in Croatian language. Data sample consists of 420 documents and 338,706 tokens. A small sample was manually tagged for terminology to be used in an initial experiment. The approach for terminology extraction is knowledge-driven and consists of differential analysis of reference and domain-specific corpora. Specific method used is log-likelihood ratio test. Experiment deals with different reference corpora and linguistic pre-processing. First results are promising. Further research guidelines are discussed.

Key words: terminology extraction, data sample, log-likelihood ratio test

Introduction

Text mining deals with information detection in natural language texts. One area of text mining is called terminology extraction which applies to (semi)automatic extraction of technical terms of a specific domain. A list of

technical terms is a requirement for e.g. specialized dictionaries. This list can help to understand an area of expertise. There are two methods to approach this problem: statistics and linguistics. Statistical method is concerned with the idea of differential analysis, which is to find a correlation between specialized lexicon and general lexicon. Linguistic methods process the text mostly on morphological and syntactic level finding proper term candidates. Hybrid methods combine these two approaches (Witschel, 2004).

Building a data sample

Documents collected for Synopsis corpus were downloaded from official web pages of the Faculty of Humanities and Social Sciences (Online arhiva dokumenata, 2009). Documents contained 420 synopses of doctoral theses accepted in the period from 2004 to 2009. They were exclusively digital texts in .doc format with a mostly uniform structure, which made it easier to import it into a database. Importing was done manually into a database management system Access 2003.

Table 1 shows the elements of the synopses with the number of synopses not containing the specific element.

Table 1: Elements of synopses

Name of element	Number of synopses with empty part (%)
title	0 (0.00%)
introduction	325 (77.38%)
theoretical background	1 (0.02%)
narrower field of work	1 (0.02%)
aims and problems of research	1 (0.02%)
methodology	6 (1.43%)
expected scientific and/or practical contribution	17 (4.05%)
structure of thesis	326 (77.62%)

Processing

After importing data into the database, Synopsis corpus was verticalized i.e. tokenized. The token rule states that a token is a constant array of letter characters, wherewith the digits and punctuation are eliminated.

Synopsis corpus was semi-automatically lemmatized. Using several specialized databases helped detect a number of tokens and matching lemmas with its word category, while the rest was lemmatized by hand. Databases used for lemmatization are following: lexical database of the Croatian literary language (Kržak, 1985), Croatian Frequency Dictionary (Moguš, 1999), a database of surnames (Boras, 2003) and a database of settlements.

Finally, by tokenizing and lemmatizing, two new columns were added to the Synopsis corpus: lemma of a particular word and its word category.

Corpus analysis

Synopsis corpus comprises 420 synopses of doctoral theses at the Faculty of Humanities and Social Sciences, University of Zagreb, accepted in the period from 2004 to 2009 written mostly in Croatian language. 305 synopses fall under the field of humanities (72.62%), while the rest of 115 fall under the field of social sciences.

Corpus has 338,706 tokens, of which 98.84% (334,799) are written in Croatian, while the rest of 1.16% (3,907) is written in other languages[1].The average size is 806.44 of tokens per document.

Corpus has 45,788 types, of which 95.08% are Croatian, while the rest of 4.92% (2,254) are in other languages. The average number of types per document is 51.32.

In Synopsis corpus one can find 338,706 tokens and 45,788 types, which makes a type-token ratio of 0.135. Researching on a corpus consisting of documents from the field of finances, (Tadić, 2003) detected that the type-token ratio for that corpus is 0.05. Comparing it to Croatian Frequency Dictionary (Moguš, 1999) where it is 0.119, they gave a possible explanation of why it is unusually high: "… the vocabulary in the field of finances shows less variation in inflection as well as limited number of different lexical entries than the general vocabulary" (Tadić, 2003). If we consider that argument to be true, the opposite statement would be an explanation of why type-token ratio for Synopsis corpus is lower than the one of Croatian Frequency Dictionary. This should not be a surprise if we keep in mind various subfields of humanities and social sciences (Table 2).

Table 2: Subfields of humanities and social sciences

Field of humanities	Field of social sciences
Philosophy	Political science
Philology	Information sciences
History	Sociology
Art history	Psychology
Science of art	Science of education
Archaeology	
Ethnology and anthropology	

Initial experiment

The idea behind the initial experiment is to get a feel for the data and the terminology extraction problem in general.

[1] Languages other than Croatian that can be found in Synopsis corpus: English, Latin, German, Italian, French, Portuguese, Hungarian, Slovenian, Czech, Polish, Serbian, Romanian, Slovak, Greek, Old English, Dutch, Ikavian Croatian, Spanish, Istro-Romanian, Middle High German, Bosnian, Kajkavian Croatian, Turkish and Swedish.

The sample which was manually tagged and used as a gold standard is rather small. It consists of only one article which has 671 tokens. The sample is tagged by only one person so no interannotator agreement can be computed. There is also just this small tagged sample meaning that there is no possibility of having a development and an additional testing corpus which would make the methodology more accurate.
The sample was tagged in a straightforward fashion - the sample is verticalised and the rows containing a terminus or part of a multiword terminus are given an additional column with the value 1. Other tokens are given the value 0. Since in the corpus preprocessing lemmatization and part-of-speech tagging are performed, this information is also provided in the sample in the form of two additional columns.
The frequency of specific syntactic patterns is shown in Table 3. The data shows that most frequent patterns are the simple ones. Nevertheless, in such a small sample highly complex patterns also occur. One example showing very clearly the syntactic complexity of the text is the following: “... postmodernom ili postindustrijskom, a kod nas i postsocijalističkom društvu.” This phrase contains actually three terms: “postmoderno društvo”, “postindustrijsko društvo” and “postsocijalističko društvo”. It is very common in the whole sample that more terms share a common head in the noun phrase. Because of this syntactic complexity, in this experiment we will try to locate only tokens that are terms or just part of terms, and not their whole phrases. One of the obvious reasons for this is the lack of syntactic language tools for Croatian language.

Table 3: Frequency of specific syntactic patterns in tagged sample (N – noun, A – adjective, C – conjunction, x – not part of the term, N(g) – noun in genitive form, A(g) – adjective in genitive form)

syntactic pattern	**frequency**
N	11
AN	8
A	4
NA(g)N(g)	3
ANCN	3
ACAN	2
AxAxxxxAN	1
NN(g)CN(g)	1
NN(g)	1

The method used to identify tokens that are possible termini or parts of multiword termini is the log-likelihood ratio test introduced by Dunning (Dunning, 1993). This method is chosen as the first to be experimented on because of its popularity in the differential analysis community (Kiss, 2002; Witschel, 2005; Kuhn, 2009). The log-likelihood ratio compares two statistical hypotheses - the zero hypothesis that the token distribution in the corpus of interest and a well

balanced reference corpus is the same, and the alternative hypothesis - that they are not. In the Dunning log-likelihood ratio test the binomial distribution is used. The binomial likelihood of a token is computed as

$$L(p,k,n) = p^k (1-p)^{n-k}$$

with

$$p = \frac{k}{n}$$

where k is the token frequency and n the size of the corpus. The logarithm of the likelihood ratio is computed as

$$-2\log\lambda = 2[\log L(p_1,k_1,n_1) + \log L(p_2,k_2,n_2) - \log L(p,k_1,n_1) - \log L(p,k_2,n_2)]$$

with

$$p = \frac{k_1 + k_2}{n_1 + n_2}$$

The higher the -2logλ the more significant is the difference between the term frequencies. If p_1 is greater than p_2 for a specific token, than the -2logλ value shows how more important the token is for the first corpus and vice versa.
There are three different reference corpora used in the research - the corpus described in this paper and a small and a large newspaper corpus.
The large newspaper corpus is built from the on-line version of the daily newspaper Vjesnik. The initial size of the corpus is 746,683 tokens. Numerals and interpunctions are not included in the corpus. The corpus is also verticalized and additional information like the lemma and part-of-speech are added as separate columns. This reference corpus was morphosintactically tagged in a different manner than the corpus described in this paper. A trigram statistical tagger (Agić, 2006) was used and no additional human intervention was undertaken. This reference corpus is called “Vjesnik1”.
The small newspaper corpus is just a subset of the large newspaper corpus. It consists of 70,000 tokens. In this research it is called “Vjesnik2”.
The corpus described in this paper used as a reference corpus consists of 338,035 tokens. It includes all documents but the one used as the gold standard. This corpus is also verticalized and lemma and part-of-speech information is also present. This reference corpus is called “Synopsis”.

The possible advantage of this reference corpus to the Vjesnik reference corpora might be that the lemmatization and POS tagging method was identical as in the gold standard. The disadvantage could be the non-representativeness of this corpus. The newspaper corpus is also not really a well-balanced reference corpus, but is probably nearer to that idea than this one.
The free parameter that has to be optimized when using the log-likelihood ratio test is the result of the test. The -2logλ value will be optimized concerning evaluation measures computed by comparing the gold standard and the result the method produces. Normally, an additional free parameter would be the minimum frequency of a token, but in this experiment this parameter will be fixed to 1. One of the arguments for doing so is the small size of the sample the experiments are performed on.
Three evaluation measures are computed on the classical measures of precision and recall - $F_{0.5}$, F_1 and F_2. The parameter optimization is performed concerning the F_2 measure. The reason for that is the most frequent usage of terminology extraction methods. Mostly the output is given to human specialists and therefore recall is more important than precision.
In this experiment baselines are considered random results. This means that when identifying terminology in the source without any POS-filtering, the probability of finding a terminus randomly is 70 divided by 671, i.e. 10.43%. This also means that on average every tenth token is a terminus.
The first experiment uses all three reference corpora. As features it uses plain lowercase tokens. In all cases the -2logλ is optimized concerning the F_2 measure. In all experiments the -2logλ measure takes values in range from 1 to 15 with step 1. The baseline is 0.104. The results are shown in Table 4.

Table 4: Evaluation measures regarding the reference corpus (RC) when using tokens as features

RC	precision	recall	$F_{0.5}$	F_1	F_2	-2logλ
Vjesnik1	0.183	0.757	0.215	0.294	0.465	7
Vjesnik2	**0.194**	**0.757**	0.228	0.309	**0.479**	7
Synopsis	0.180	0.743	0.212	0.290	0.457	4

The different experiment layouts show pretty similar results. The only significant difference is the -2logλ optimal measure. When using any version of the Vjesnik reference corpus, it is 7 and, when using the Synopsis corpus, it is 4. The reason for that is probably the greater similarity between the gold standard and the Synopsis reference corpus. Interesting is also that the smaller newspaper reference corpus did not lower the result; on the contrary, it improved it, but not significantly. The reason for that can, of course, be also pure coincidence, i.e. the content of the smaller corpus.
In general all reference corpora show a significant improvement in comparison to the random baseline.

The distribution of part of speech in the gold standard and the optimal result in the previous experiment (Vjesnik2 as reference corpus and -2logλ=7) is shown in Table 5. As expected, the gold standard consists only of nouns and adjectives with exception of the conjunction "i" ("and"), since this conjunction is used where terms share the same head and human annotator considered it part of the multi-term noun phrase. The fact that the result consists also of other parts of speech (especially verbs) indicates the potential usefulness of a POS filter that will be introduced later in the experiment.

Table 5: Distribution of part of speech in the gold standard and the optimal result in the first experiment

part of speech	**gold standard**		**result**	
	type	token	type	token
noun	35	39	97	144
adjective	22	25	71	84
verb	0	0	17	27
conjunction	1	6	0	0
pronoun	0	0	2	6
number	0	0	3	3
abbreviation	0	0	3	3

The second experiment has a similar layout to the first experiment, it just uses lemmata as features and not tokens. The baseline of this experiment is the same as in the previous case 0.104. The results are shown in Table 6.

Table 6: Evaluation measures regarding the reference corpus (RC) when using lemmata as features

RC	**precision**	**recall**	**$F_{0.5}$**	**F_1**	**F_2**	**-2logλ**
Vjesnik1	0.118	0.514	0.139	0.191	0.307	1
Vjesnik2	**0.125**	**0.600**	0.148	0.206	**0.340**	1
Synopsis	0.152	0.486	0.176	0.231	0.337	2

The data show a rather surprising result – a decline in all three reference corpora when using lemmata as features in comparison to using tokens. A possible explanation is that morphological normalization added less information than it was written in specific word forms. Interestingly, the smaller newspaper reference corpus secures a tight win in this experiment again. Second best is the synopsis corpus. The reason for that is probably the fact that lemmatization in the Synopsis reference corpus and the sample was realized with the same method while the Vjesnik corpus was lemmatized by a different method using different language resources. Optimal -2logλ is in all cases very low. The reason for that is the unification done by lemmatization, namely the number of different values in the sample is now much lower.

The third experiment introduces a POS filter. Namely, only nouns and adjectives are allowed as results. This method should improve the results since al-

most all termini are or consist of only adjectives and nouns. The random baseline for this experiment is higher since now candidate termini are only nouns and adjectives That means that only 253 nouns and 134 adjectives, ie. 387 tokens are termini candidates. The probability of picking a terminus on random is 70/387, ie. 18.1%. The results are shown in Table 7.

Table 7: Evaluation measures regarding the reference corpus (RC) when using a POS filter and tokens as features

RC	precision	recall	$F_{0.5}$	F_1	F_2	-2logλ
Vjesnik1	0.220	0.813	0.258	0.347	0.528	7
Vjesnik2	**0.205**	**0.891**	0.242	0.333	**0.534**	5
Synopsis	0.211	0.859	0.248	0.338	0.532	3

These results show, as presumed, a significant improvement in comparison to the previous methods. Again, the winner is the Vjesnik2 reference corpus. In this method, the -2logλ is slightly lower than when not applying a POS filter. Interestingly, the improvement of the POS filter is not too big. The reason is that the log-likelihood ratio test does a pretty good job in identifying primarily nouns and adjectives. The presumption is that the distribution of other part-of-speech entities is rather constant. In Table 8 the distribution of part of speech of the optimal results of the first experiment in comparison to the distribution on the whole reference corpus.

Table 8: Comparison of POS distributions in the result of the first experiment and the Vjesnik1 reference corpus

part of speech	reference corpus	result	difference
noun	0.384	0.56	+45.8%
adjective	0.274	0.32	+16.8%
verb	0.101	0.10	-1.0%
other	0.242	0.02	-91.7%

The results show that amost 90% of the tokens in the result are nouns and adjectives. In the newspaper reference corpus they make some 55% of all the tokens. Verbs are rather constant. Nouns and adjectives gain in the probability mass from other parts of speech. The conclusion is that other parts of speech are equally distributed over different samples. Nouns are mostly differently distributed. Adjectives take the second place. Verbs do not show any difference in the probability mass.

Further research

Further research will include a bigger tagged sample. This sample, namely, contains only 671 tokens.

Different document sizes will be included in the research. For differential analysis the length of the domain-specific corpus is of great importance.

Experimenting with different text complexity will also be of interest. The doctoral synopses texts are very complex which was shown by the high type-token ratio. This sample is especially syntactically complex. That fact would make the process of finding syntactic cues for termini identification very hard. Samples will also be annotated by more annotators. That will provide us with the measure of interannotator agreement.
In further research the methodology of using distinct development and testing samples will be followed.
Further experiments will be conducted concerning the size and content of reference corpora.
The minimum frequency criterion for document features will also be included.
More methods of differential analysis will also be experimented with.

Conclusion

This paper describes the process of building a data sample for terminology extraction and an initial research on the data.
The data sample consists of 420 documents and 338,706 tokens. The type-token ratio is high which indicates complex vocabulary. The sample is syntactically particularly complex.
At this point just a small portion of the sample is tagged. This part of the sample is used as a gold standard for the initial research.
An interesting result of the research is that a smaller newspaper reference corpus yields better results than the two other corpora. Additional research is necessary to inspect the reasons for such results.
When using lemmata as document features, results were consistently worse. We assume that more information was lost by not including tokens than information was gained by including lemmata. A combination of both features could further improve results.
The POS filter improves the results significantly by choosing only nouns and adjectives as candidate termini. When not using the POS filter, nouns and adjectives are chosen more often than by chance. This leads to the conclusion that they differ between corpora more than verbs and, especially, other parts of speech. Nouns differ more than adjectives.
In general, the investigated methods achieve significantly better results than the random baseline.
Further research will include a bigger and more versatile gold standard, different reference corpora, more annotators and a more complex methodology.

References

Agić, Željko; Tadić, Marko. Evaluating Morphosyntactic Tagging of Croatian Texts. // *Proceedings of the 5th International Conference on Language Resources and Evaluation* / Genova: ELRA, 2006.

Boras, Damir; Mikelić, Nives; Lauc, Davor. Leksička flektivna baza podataka hrvatskih imena i prezimena. // *Modeli znanja i obrada prirodnoga jezika.* / Zagreb: Zavod za informacijske znanosti, 2003, 219-237

Dunning, Ted. Accurate Methods for the Statistics of Surprise and Coincidence. // *Computational Linguistics.* 10 (1993), 1; 61-74

Kiss, Tibor; Strunk, Jan. Scaled log likelihood ratios for the detection of abbreviations in text corpora. // *Proceedings of the 19th International Conference on Computational Linguistics* / ACL, 2002, 1-5

Kržak, Miroslav; Boras, Damir. Rječnička baza hrvatskog književnog jezika = Lexical Data Base of the Croatian Literary Language. // *Informatologia Yugoslavica*. 17 (1985), 3 4; 223-242.

Kuhm, Adrian. Automatic labeling of software components and their evolution using log-likelihood ratio of word frequencies in source code. // *Mining Software Repositories, 2009. MSR '09. 6th IEEE International Working Conference on* /2009, 175-178

Moguš, Milan; Bratanić, Maja; Tadić, Marko: Hrvatski čestotni rječnik. Zagreb : Školska knjiga, Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, 1999.

Online arhiva dokumenata. 17.07.2009. http://www.ffzg.hr/dokument/index.php?cid=1801 (20.07.2009.)

Tadić, Marko; Šojat, Krešimir. Finding Multiword Term Candidates in Croatian. // *Proceedings of Information Extraction for Slavic Languages 2003 Workshop* / Borovets : BAS, 2003, 102-107.

Witschel, Hans Friedrich. Terminology Extraction and Automatic Indexing -- Comparison and Qualitative Evaluation of Methods. // *Proceedings of Terminology and Knowledge Engineering (TKE)* / 2005

Witschel, Hans Friedrich. Text, Wörter, Morpheme – Möglichkeiten einer automatischen Terminologie-Extraction. (Diploma thesis) Leipzig, 2004.