Recognition of non-domain phrases in automatically extracted lists of terms

Agnieszka Mykowiecka
Institute of Computer Science PAS
Jana Kazimierza 5
01-248 Warsaw, Poland
agn@ipipan.waw.pl

Małgorzata Marciniak
Institute of Computer Science PAS
Jana Kazimierza 5
01-248 Warsaw, Poland
mm@ipipan.waw.pl

Piotr Rychlik
Institute of Computer Science PAS
Jana Kazimierza 5
01-248 Warsaw, Poland
rychlik@ipipan.waw.pl

Abstract

In the paper, we address the problem of recognition of non-domain phrases in terminology lists obtained with an automatic term extraction tool. We focus on identification of multi-word phrases that are general terms and discourse function expressions. We tested several methods based on domain corpora comparison and a method based on contexts of phrases identified in a large corpus of general language. We compared the results of the methods to manual annotation. The results show that the task is quite hard as the inter-annotator agreement is low. Several tested methods achieved similar overall results, although the phrase ordering varied between methods. The most successful method with the precision about 0.75 at the half of the tested list was the context based method using a modified contextual diversity coefficient.

1 Introduction

Automatic term recognition (ATR) can be applied to achieve concept names which might be included in a domain ontology. However, lists of terms obtained in this way should be filtered to exclude terms belonging to different specialized domains which occurred within the text only by coincidence (e.g. citations); terms which are general, such as *low level* used in many different domains; and discourse markers like *point of view*. It is difficult to consider that phrases such as *low level* or *left side* are domain specific, but they play an important role in several domains, e.g. medicine or technology. Phrases like *turning point* or *difficult question* should be excluded from terminology lists. While identification of domain terms has been addressed by several researchers, the problem of general terms identification has not been studied greatly, although it poses a much harder task to cope with. We propose identifying such phrases and building a separate resource to be combined with other domain specific ontologies.

The filtering out-of-domain terms has been the subject of several studies. Most typical approaches are described in (Schäfer et al., 2015), other attempts include (Navigli and Velardi, 2004)) or (Lopes et al., 2016). Discrimination of in- and out-of-domain terms was based on identifying terms occurring more frequently in the given domain related data than in other corpora. Most of these approaches looked for terms which are more salient in particular corpora than in others and work relatively well for selecting specialized terms. In this paper we focused our attention on terms which are nearly equally frequent in many corpora and thus are hard to classify either as domain specific or general. We decided to focus on multi-word terms as most of them are not present in general wordnet-type datasets. They are also easier to classify as either domain specific or general. Thus, the evaluation of the proposed methods is more reliable.

2 Terminology extraction

We used the TermoPL program (Marciniak et al., 2016) for the ATR task. It consists of standard phases of candidate selection and ordering. TermoPL accepts morphosyntactically analyzed texts and calculates the C-value (Frantzi et al., 2000) for phrases recognized using either a built-in or customized grammar. The ATR based on the C-value coefficient allows extraction of one-word and multi-word phrases, as part

Proceedings of the 5th International Workshop on Computational Terminology,
pages 12–20, Osaka, Japan, December 12 2016.

of one common terminology list, and creates a ranked list of these terms. It allows us to compare such a list with another list obtained using the same method from a different corpus. For common terms, the program indicates for which corpora they are more representative.

In our experiments, we used a standard built-in grammar for candidate selection. It applies a simple shallow grammar describing most typical Polish noun phrases, i.e. nouns, nouns modified with adjectives placed before or after a noun (it respects case, gender and number) and nominal phrases post-modified with nominal phrases in the genitive. The ordering is performed using the slightly modified C-value coefficient. This coefficient is computed on the basis of the number of times a phrase occurs within the text, its length, and the number of different contexts this phrase occurs within the text. The definition of the C-value coefficient is given in (1).

$$\textit{C-value}(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)), & if\ r(LP) > 0, \\ l(p) * freq(p), & if\ r(LP) = 0 \end{cases} \quad (1)$$

p is a phrase under consideration,
LP is a set of phrases containing p,
r(LP) is the number of different phrases in *LP*,
$l(p) = log_2(length(p))$.

In this paper, we focus on the further stage of processing the term list, i.e. its filtering, independently of the extraction method used to obtain it.

3 Domain corpora

In our work, we analyzed six different sets of texts. The first five are domain corpora, while the last one is more general:

- *ChH* – a set of patients records from a children hospital,
- *Music* – a part of the ART Corpus[1] related to music and its history,
- *HS* – books and articles on the history of art, a part of the ART Corpus,
- *Lit* – literature papers from the ART Corpus,
- *wikiE* – a part of Polish Wikipedia with articles related to economy (`http://zil.ipipan.waw.pl/plWikiEcono`),
- *KS* – journalistic books from the Polish National Corpus (NKJP) (`http://clip.ipipan.waw.pl/NationalCorpusOfPolish`).

The details about the size of each corpus and the number of recognized terms are given in Table 1. We observed that although the total number of multi-word terms constitute about one third of all term occurrences, the number of different phrases is much higher than one half of all of them.

Table 1: Corpora statistics

corpus	tokens	#terms	#mw-terms
ChH	1,966K	26K	21K
Music	1,075K	94K	65K
HS	1,438K	157	126K
Lit	2,410K	220K	185K
wikiE	456K	57K	49K
kS	3,204K	164K	137K

Table 2 presents numbers of common multi-word nominal phrases which occurred in at least three corpora.

4 Term selection based on domain corpora

The lists of terms obtained by any ATR tool contain a large number of valid terminological expressions, but they also contain some out of domain, general and even improperly structured phrases. It had already

[1]The data will be soon available.

Table 2: Common multi-word terms

#corpora	6	5	4	3
#shared mwterms	44	353	1441	5113

been proposed to eliminate such terms using corpora-comparing log-likelihood (Rayson and Garside, 2000), Contrastive Selection via Heads (Basili et al., 2001) and Term Frequency Inverse Term Frequency (TFITF) (Bonin et al., 2010), but all these methods perform relatively well only when both corpora – domain and general – are voluminous enough. For specialized domains, we frequently do not have enough data to judge on the basis of one comparison. To make the decisions more reliable, we compare several (not necessary very big) corpora to gain the necessary information out of many comparisons. We analyze three different solutions to this problem and compare them on the same set of corpora.

I. Co-occurrence in multiple corpora The simplest approach for detecting general (or out-of-domain) terms could be identification of terms which occur in more than one terminology list. Although multi-word terms do not occur very frequently, general phrases should occur in many different contexts i.e. their frequencies could be sufficiently high. To test this hypothesis we check multi-word phrases which occur in more than three out of six tested corpora. The problem with this approach is the fact that if we decide to stick to terms which occur in all but one corpora, we may identify a small group of phrases. As for the less frequent terms, we quickly get much less reliable candidates. The second issue is that we treat equally terms that occur very frequently and those which are very rare.

II, IIa C-value standard deviation based weighting In the second method we utilize information about the strength of a particular term within each corpora, i.e. its C-value. We normalize the C-value to have the same overall sum in all corpora and assign each term a weight depending on whether it is not present in a corpus (-1), has a C-value near 0 (0.5), below 1 (1), below a selected threshold equal to 8 (2) and above it (3). Then, we count the standard deviation between all weights and order terms according to their ascending value. The top terms are equally important (or unimportant) in all corpora. Terms which only have a high C-value on some of the term lists are moved towards the end of the final ranking. This method promotes terms which are important and their relative position from the top of the list is similar. In the modified version of the method, named IIa, we used $\log_{10}$ of the C-values instead of the rigid weights (still -1 was assigned to non-present terms).

$$C_{IIa}(t) = \frac{\Sigma_{all\_corpora}\ \sigma_{\log_{10}(C\text{-}value\text{-}norm(t))}}{number\text{-}of\text{-}corpora}$$

III Another method is based on the observation made in (Lopes et al., 2016) where it is suggested that terms that appear in the contrasting corpora should have been penalized proportionally to the number of their occurrences. Thus, the absolute frequency of the term in the domain corpus is divided by a geometric composition of its absolute frequency in each of the contrasting corpora. We adapted this idea to calculate a list of general terms ordered by a geometric composition of their C-values in all the corpora examined. The higher the coefficient C_{III}, the lower the probability that the term is domain related.

$$C_{III}(t) = \prod_{\forall corporaC} (1 + \log_{10}(C\text{-}value^{C}(t))$$

II+III, IIa+III Second order methods. When analyzing the results obtained by all the above methods, we observed that the number of common terms on top of the lists computed by the II (IIa) and the III method are the smallest. Thus, we combined these two methods in one by means of linear combination of their normalized values. As the coefficients obtained by the methods are ordered in the opposite way, the equation looks as below, where α is a number between 0 and 1.

$$C_{IIa_{+}III}(t) = \alpha(1 - C_{IIa\text{-}norm}(t)) + (1 - \alpha) * C_{III\text{-}norm}(t)$$

5 Term selection based on term contexts in a general corpus

We decided to compare the results obtained with the methods described in Section 4 to a method which judges the term generality on data obtained from a single (many domain or general) corpus. This method is based on the observation that domain terms usually occur together with other terms from the same domain so their contexts mainly consist of in-domain expressions/words together with the general ones. On the contrary, general terms and functional expressions can accompany expressions from many unrelated domains and, thus, they tend to have much more diverse contexts. To measure this diversity, we apply a clustering coefficient described in (Hamilton et al., 2016) to measure a word's contextual diversity and, thus, polysemy. In method IV, we ordered all terms according to the increasing diversity coefficient $d(w)$. This coefficient measures the percentage of related context pairs within the set of pairs of contexts which are highly related to the analyzed term. A related pair of words is defined as a pair which has a non-zero Positive Pointwise Mutual Information (PPMI) value. A pair consists of two context words in the first case, and of a term and a context word in the latter.

$$d(w) = \frac{\Sigma_{c_i,c_j \in N_{PPMI}(w)} C_{N\_PPMI}(c_i, c_j)}{|N_{PPMI}(w)|(|N_{PPMI}(w)| - 1)}$$

$C_w = \{w_i : w_i \textit{ is in a context of } w\}, N_{PPMI}(w) = \{w_j \in C_w : PPMI(w, w_j) > 0\}$ and $C_{N\_PPMI}(c_i, c_j) = \{1 \text{ if } PPMI(c_i, c_j) > 0 \text{ and 0, otherwise}\}$. The PPMI value represents the strength of correlation between two words. The larger is the number of common occurrences in a relation to all possible two word pairs, the stronger correlation.

$$PPMI(w, z) = max\{log(p(w, z)/(p(w) * p(z)), 0\}$$

The tested hypothesis was whether the increasing order of this coefficient, which is aimed at reflecting the decreasing polysemy factor, represents satisfactorily the difference between the general terms that can be used in very different contexts, thus gaining different meaning, and domain related terms which are less polysemous. As in principle, a general term could not have any highly related contexts, we suggest modifying the $d(w)$ coefficient by replacing the nominator by the number of all possible context pairs (limiting the context only by the number of occurrences not by a non-zero PPMI). The modified d_M coefficient is defined as follows:

$$d_M(w) = \frac{\Sigma_{c_i,c_j \in N_{PPMI}(w)} C_{N\_PPMI}(c_i, c_j)}{|C_w|(|C_w| - 1)}$$

To deal with small corpora, for which the original method is unable to judge many terms as they do not have any contexts classified as related, a variant of method IV is introduced. For such a case, we propose an additional step for selecting terms which are similar to the analyzed one. Similarity is defined here as the cosine similarity of the vectors from the word2vec model (Mikolov et al., 2013) trained on the corpus in which multi-word term occurrences were replaced by the concatenation of the term elements and thus were treated as singular model features. We trained the standard continuous bag-of-words model with the 5 word window and 200 features. Next, we combined all the contexts of a term with the contexts of all terms for which the similarity was greater than 0.44. We observed that, for multi-word terms, the similarity coefficient is generally lower than for one-word terms and that, in small corpus, the higher threshold provides very few similar terms. In Tables 3–5, we gave examples of similar multi-word terms calculated on the basis of the domain corpora described in Section 3. For the first two expressions, the method found helpful similar terms, while Table 5 rather contains terms unrelated to the considered one, i.e., *dzieło stworzenia* 'act of creation'.

In the next step, we used the same procedure as before, that is we counted the $d(w)$ diversity coefficient for all contexts of similar terms clustered together.

Table 3: Similar multi-word terms for *duże wrażenie* 'big impression'

term	similarity	translation
ogromne wrażenie	0.755	'huge impression'
wielkie wrażenie	0.740	'great impression'
dobre wrażenie	0.514	'good impression'
wielki wpływ	0.463	'great influence'

Table 4: Similar multi-word terms for *dziwiętnasty wiek* 'nineteenth century'

term	similarity	translation
XVII wiek	0.506	'17th century'
XIX wiek	0.503	'19th century'
XVIII wiek	0.497	'18th century'
XX wiek	0.489	'20th century'
wiek XVIII	0.487	'18th century'
dwudziesty wiek	0.483	'twentieth century'
początek xx wiek	0.448	'beginning of the twentieth century'
XIX stulecie	0.448	'19th century'
wiek dziewiętnasty	0.438	'nineteenth century'
początek wieku	0.438	'beginning of the century'
minione stulecie	0.434	'past century'

6 Evaluation

To evaluate our method we prepared two manually annotated lists. The first one, called *COM*, consists of 7151 terms which occur in at least three of the six selected corpora. Annotation was done by two annotators and then the third one resolved the conflicts to obtain the gold standard annotation (GS). The annotators introduced five labels representing *non-terms, general-terms, domain-terms-used-generally, domain-terms, improper-phrases*. At the evaluation stage as *general-terms* we treated the first three classes together. Table 6 includes the number of annotations of each type. The difficulty of the task and the lack of the strict guidelines is reflected in a relatively low Cohen's kappa-coefficient which is equal to 0.45. As the first test set contained a lot of phrases located very low on the ranked terminological lists, we also prepared the second test set (*MFQ*) to verify our context based method. This test set is based on the first 1000 terms from the terminological lists obtained separately for all corpora except the medical one.[2] The resulting 3250 terms were annotated by the same two annotators. To reduce the influence of the subjectivity of judgments (the kappa coefficient was 0.5), the final test set contains only 2341 terms which were annotated identically by both annotators. 964 terms are included in both test sets.

As our results are ranked lists, we had to introduce a threshold indicating which part of the lists should be treated as general terms. For the first method, we selected terms which occur in at least 4 corpora; for the others, we treated 70% of the lists as general terms. This is roughly the most desirable partition as the *COM* test set contains a little more than 73% of general terms.

Table 7 gives the number of common annotations made using the above methods and the threshold.

For the evaluation of the IV method we performed the experiments in which we used two data sets and two lists of terms. The first (*art*) corpus consisted of four of the corpora described in section 3 (all except the hospital data set – ChH). It consists of about 845K tokens. The second data set ((*nkjp+art*) is much larger, with 1.3G words from the complete NKJP — National Corpus of Polish Language (Przepiórkowski et al., 2012) added to the (*art*) corpus. The term list is the same list of 7151 terms described above. While counting the diversity coefficient $d(w)$ we only selected contexts which were

[2]Most terms from this set of data occur very frequently in the NKJP corpus.

Table 5: Similar multi-word terms for *dzieło stworzenia* 'act of creation'

term	similarity	translation
kłos zboża	0.459	'ear of grain'
postać ludzka	0.439	'human figure'
świat widzialny	0.438	'visible world'
wspólne dzieło	0.431	'joined act'

Table 6: Manual annotation

	COM test set			*MFQ* test set		
	An1	An2	GS	An1	An2	GS
general-term	6228	5228	5273	1493	1296	999
non-general-term	799	1641	1741	1571	1893	1342
error	124	282	237	175	51	–

strings containing only lower case letters. We excluded named entities from this set. We also disregarded the most common words (e.g. prepositions and pronouns). For this purpose, we used the list of stop words from the Wikipedia page. As the PPMI value is biased towards low frequency phenomena, we took into account only pairs which occur in NKJP more than 5 times.

Table 7: Common annotations for *COM* test set

method	I	II	IIa	III	IIa+III	IV_{art}	$IV_{nkjp+art}$
GS	2970	3720	3717	3726	3187	4020	4762
I	-	3818	3752	5229	4167	2791	2983
II	-	-	6100	5722	6252	2285	3411
IIa	-	-	-	1888	6696	2364	3387
III	-	-	-	-	2301	3646	3394
IIa+III	-	-	-	-	-	2532	3413
IV_{art}	-	-	-	-	-	-	3772

For all methods we counted how many terms annotated as general in the GS file were found in each part of the ranked lists. The results for every 500 element segments are shown in Figure 1, while Figure 2 shows the overall precision by steps of 500 terms.

Figures 1 and 2 show that the most methods do not differ much. The most stable results were achieved for IIa and the combination IIa+III. For the latter method we tested several values of α from 0.2 to 0.8 and the best results were obtained with α 0.4. The methods I and III are shown to be the least consistent. The method IV showed the quickest decrease of the percentage of the general terms for each five hundred positions, thus proving to be the most selective one.

In the second experiment, in which we check the contexts of the phrases, the results obtained for a small corpus containing four sets described in Section 3 (IV_{art}) turned out to be rather poor. The list of terms with non-zero related contexts was very short — it contained only 301 elements. The resulting precision was only 0.33. For this data set, the addition of similar terms ($IV_{art^{add}}$) improved the results. In this approach we found relevant contexts for 948 terms with a precision equal to 0.64 for the first 500 elements and 0.5 for the entire set. For the big corpus, the results achieved by adding similar terms ($IV_{nkjp+art^{add}}$) were slightly worse, as was expected. Table 8 summarizes the results and presents the precision obtained by all our methods for the first 500 elements and for the entire set (* indicates that the method did not process the entire *COM* list).

In the next set of experiments we tested more extensively different variants of the IV method which is based on contextual information. On two term test sets described above, apart form the basic version of the method, we tested the newly introduced d_M coefficient and the non-uniform treatment of the context words. In a weighted d^w schema we assigned smaller weights to context words which are more distant from the given term (in a 5 word window, the farthest word has weight equal to 0.2 while the closest neighbour has the weight of 1). We performed tests on the big *nkjp+art* corpus. The results shown in

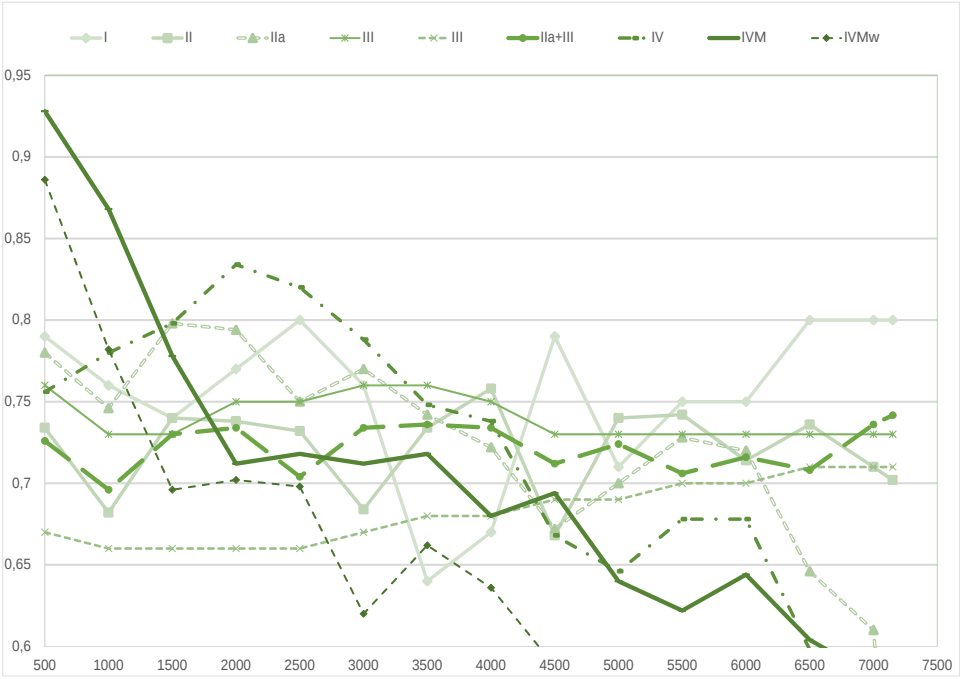

Figure 1: Percentage of general terms for every 500 terms individually for all methods – *COM* test set

Table 8: Precision of all the methods – *COM* test set

	I	II	IIa	III	IIa+III	IV_{art}	$\text{IV}_{art^{add}}$	$\text{IV}_{nkjp+art}$	$\text{IV}_{nkjp+art^{add}}$
first 500 terms	0.79	0.73	0.78	0.73	0.73	0.33	0.64	0.74	0.66
entire list	0.47	0.58	0.62	0.58	0.58	*0.33	*0.50	0.58	0.69

Table 9 confirm improvement in cases where the d_M coefficient was used. The number of the general terms at the beginning of the list is higher and this proportion constantly decreases, which was not the case for the other methods. The non-uniform weighting of context words caused deterioration of results.

Table 10 shows how many terms were filtered out from the top part of terms in the 5 domain corpora. We tested lists of, at most 1800, top general terms obtained by 9 methods separately. We tested only the top parts of all domain term lists consisting of 10K terms. It shows that method III is more efficient in eliminating phrases from the top of the term list than the other methods. Unfortunately, it concerns both types of terms: out-of-domain terms and false positive out-of-domain terms.

7 Conclusions

Differentiation between general terms and domain specific terms is a hard task. The methods proposed in this paper allows for preselecting sets of phrases containing more than seventy percent of general terms.

For the methods based on domain corpora, the most efficient and, at the same time, simple method relies on standard deviation for C-value coefficient. Such a set can help when preparing lists of concepts shared by several domains. However, its usage for the task of eliminating general terms from the terminological list obtained automatically is limited, as many of these candidates are located low on these

Table 9: Precision of different variants of IV method, *nkjp+art* corpus

	COM test set				*FRQ* test set		
	IV	IV^M	IV^{Mw}		IV	IV^M	IV^{Mw}
first 500 terms	0.74	0.93	0.89	first 250 terms	0.72	0.83	0.77
second 500 terms	0.72	0.90	0.78	second 250 terms	0.53	0.61	0.55
entire list	0.58	0.64	0.62	entire list	0.55	0.62	0.62

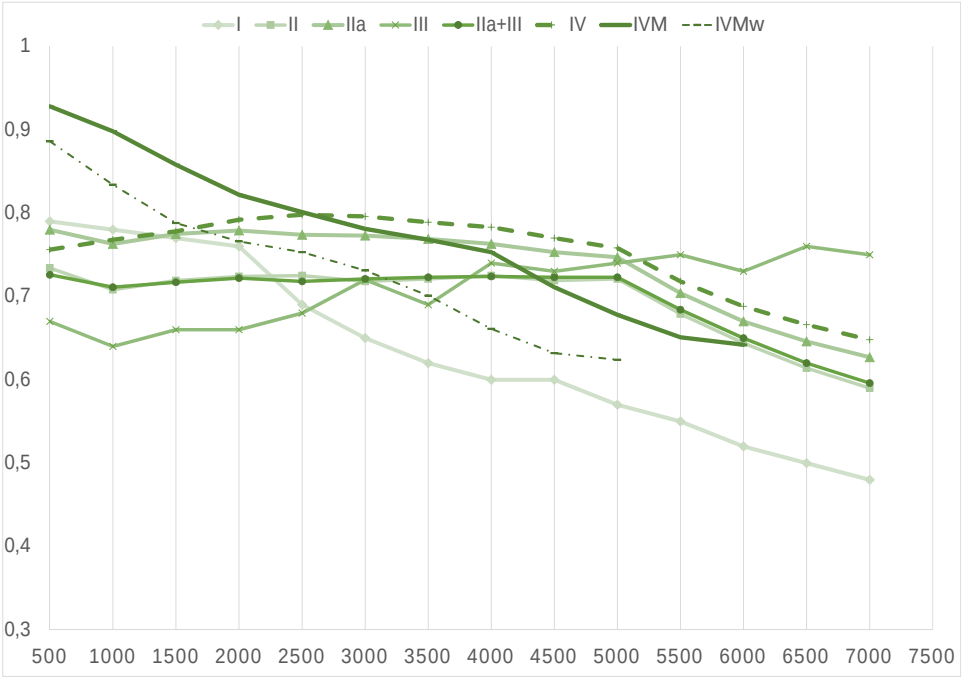

Figure 2: Precision of all methods at every 500 terms – *COM* test set

Table 10: Filtered out out-of-domain terms in 10K top terms

corpus	I	II	IIa	III	IIa+III	IV_{art}	$IV_{nkjp+art}$	$IV^{M}_{nkjp+art}$	$IV^{Mw}_{nkjp+art}$
ChH	**89**	41	43	83	52	3	61	66	77
HS	359	27	50	**482**	124	64	290	345	368
Music	387	27	71	469	145	46	**484**	449	450
Lit	640	37	86	**819**	179	79	334	740	747
wikiE	262	27	71	**301**	138	30	222	260	286

lists. The method III seems to be the best for selecting highly located general terms but it needs further research.

The method based on term contexts requires a large corpus for context recognition. The experiments performed on the small corpus gave rather poor results, but they were improved if contexts of similar terms were added. On lager corpus, this method gave much better results – the percentage of the general terms at the top of the ranked list was larger than average and larger than for all the other methods. The best variant of the method is based on the newly introduced d_M coefficient which measures the relative number of highly inter-related contexts.

Using vector similarities to expand the number of contexts did not improve results on a large corpus. For future research, we plan to use word2vec model for extending the list of general terms by phrases close to those recognized in the data as we observed many similar general terms to be relatively well clustered by cosine similarity within a model using 200 vector dimensions.

8 Acknowledgements

The paper is partially supported by the National Science Centre project *Compositional distributional semantic models for identification, discrimination and disambiguation of senses in Polish texts* number 2014/15/B/ST6/05186.

References

Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. A contrastive approach to term extraction. *Terminologie et intelligence artificielle. Rencontres*, pages 119—128.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Malta*, pages 19—21.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *Int. Journal on Digital Libraries*, 3:115–130.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Lucene Lopes, Paulo Fernandes, and Renata Vieira. 2016. Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. *Knowledge-Based Systems*, 97:237–249.

Małgorzata Marciniak, Agnieszka Mykowiecka, and Piotr Rychlik. 2016. TermoPL — a flexible tool for terminology extraction. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751.

Roberto Navigli and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2):151–179.

Adam Przepiórkowski, Mirosław Bańko, R. L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9, WCC '00*, pages 1—-6.

Johannes Schäfer, Ina Rösiger, Ulrich Heid, and Michael Dorna. 2015. Evaluating noise reduction strategies for terminology extraction. In Thierry Poibeau and Pamela Faber, editors, *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence, Universidad de Granada, Granada, Spain, November 4-6, 2015.*, volume 1495 of *CEUR Workshop Proceedings*, pages 123–131. CEUR-WS.org.