Recognition of irrelevant phrases in automatically extracted lists of domain terms

Abstract

In our paper, we address the problem of recognition of irrelevant phrases in terminology lists obtained with an automatic term extraction tool. We focus on identification of multi-word phrases that are general terms or discourse expressions. We defined several methods based on domain corpora comparison and a method based on contexts of phrases identified in a large corpus of general language. The methods were tested on Polish data. We used six domain corpora and one general corpus. Two test sets were prepared to evaluate the methods. The first one consisted of many presumable irrelevant phrases, as we selected phrases which occurred at least in three domain corpora. The second set mainly consisted of domain terms as it was composed of the top-ranked phrases automatically extracted from the analysed domain corpora.

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 a precision of about 0.75 on half of the tested list, was the context based method using a modified contextual diversity coefficient.

Although the methods were tested on Polish, they seems to be language independent.

Keywords: automatic term recognition, domain corpora, irrelevant phrases, similar phrases

1. Introduction

Automatic term recognition (ATR) can be applied to recognize concept names which might be included in a domain ontology. Unfortunately, lists of term candidates obtained in this way contain quite a lot of phrases that should not be considered as the domain terms. The lists should be filtered to exclude irrelevant phrases, e.g. 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 expressions like *point of view*.

Although the last two groups are a little different, they contain phrases that can hardly be considered as domain specific. They may, however, play an important role in several domains, e.g. medicine or technology. On the contrary, phrases like *turning point* or *difficult question* should be excluded from any terminology list.

While identification of domain terms has been addressed by several researchers, the problem of irrelevant phrases identification of irrelevant phrases 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 of out-of-domain terms has been the subject of several studies. The most typical approaches are described in (Schäfer, et al. 2015), other attempts include (Navigli and Velardi 2004) and (Lopes, Fernandes and Vieira 2016). Discrimination of in- and out-of-domain terms is based on identifying terms occurring more frequently in the given domain related data than in other corpora. Most of these approaches look for terms which are more salient in particular corpora than in others and work relatively well for selecting specialized terms.

In the paper, we test methods for selecting irrelevant phrases by comparison of more than two corpora. We focus our attention on phrases which are nearly equally frequent in many corpora, and thus are hard to classify either as domain specific or not.

We decided to deal with multi-word phrases only as most of them are not present in general WordNet-type datasets, so they need to be classified using other methods. They are also easier to classify as either domain specific or general as they are usually unambiguous. Thus, the evaluation of the proposed methods is more reliable. In our work, we process Polish texts but the methods of term selection can be applied to other languages without change.

2. Terminology extraction

In this work, for the purpose of terminology extraction we used the TermoPL program (Marciniak, Mykowiecka and Rychlik 2016). The process consists of standard phases of

candidate selection and ordering. TermoPL accepts morphosyntactically analyzed texts and calculates the C-value (Frantzi, Ananiadou and Mima 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 units and creates a ranked list of these terms.

In our experiments, we used a standard built-in grammar for candidate selection. It is a simple shallow grammar describing most typical Polish noun phrases, i.e. nouns, nouns modified with adjectives placed before or after a noun (the rules respect case, gender and number agreement) and nominal phrases post-modified with nominal phrases in the genitive. The ordering of phrases 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 is used in. The definition of the C-value coefficient is given in (1).

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

(1)

p is a phrase under consideration, *LP* is a set of contexts, i.e. phrases containing *p*, *l(p)* = $log_2(length(p))$ or 0.1 for one-word terms , *r(LP)* is the number of different phrases in *LP*. The definition of the *LP* set may vary as Frantzi et al. (2000) do not give a precise interpretation of the notion of context. The definition of the context is even more problematic in the case of free order languages, and for Polish it is discussed in (Marciniak and Mykowiecka 2015). Following the observations made there, we assume that the contexts of a phrase consist of pairs of direct left and right neighbouring words of the phrase combined together. So for the phrase: *ściany wschodniej*[1] 'eastern wall' and the following phrase containing it: *surowa cegła ściany*

[1] The example neglects the problem of cases.

wschodniej kościoła parafialnego ‘raw brick of the eastern wall of the parish church’ the context words are: *cegła,* and *kościoła* ‘brick, church’

ATR program used on one corpus extracts a list of potential terms which consists of valid (according to a chosen grammar) terms, i.e. lexical units which designate concepts in a particular domain, and irrelevant phrases which does not represent such concepts and should be filtered out. In our paper, we focus on this post-processing stage of preparing the final term list, i.e. filtering out unwanted items independently of the extraction method used to obtain it.

3. Domain corpora

We analysed 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’s hospital,
- Music – a part of the ART Corpus[2] 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 the economy (http://zil.ipipan.waw.pl/plWikiEcono),
- KS – journalistic books from the Polish National Corpus (NKJP) (http://clip.ipipan.waw.pl/NationalCorpusOfPolish).

Details about the size of each corpus and the number of terms recognized by the TermoPL tool in these texts are given in Table 1. We observed that the total number of multi-word phrases constitute about one third of all phrases occurrences, but the number of different phrases is much higher than one half of all of them.

[2] The corpus consists of texts (books chapters and journal papers) concerning fine arts.

Table 1 Corpora statistics (in thousands)

corpus	#tokens	#terms	#term occ.	#mw-terms	#mw-terms occ.
ChH	1,966	25	543	21	169
Music	1,075	93	408	63	98
HS	1,438	154	612	124	198
Lit	2,410	221	486	185	450
wikiE	456	55	221	47	78
KS	3,204	160	957	133	191
All	10,549	707	3,228	574	1,184

Table 2 shows the numbers of common multi-word nominal phrases which occurred in at least three corpora at least once or at least twice in each of them. It may be observed that imposing any frequency limit diminishes this number significantly so we did not introduce any.

Table 2 Shared multi-word phrases

#corpora	min. occ.	6	5	4	3
#shared mw-terms	1	37	318	1371	5147
	2	10	60	198	724

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 been proposed to eliminate such terms using corpora-comparing log-likelihood (LL) (Rayson and Garside 2000). This approach uses a coefficient calculated on term frequencies in two corpora. It shows to what extent the term distribution in both corpora is uniform. The higher term coefficient indicates that the term is more specific to one of the domains, but the method does not indicate to which one. Another approach uses a slightly modified TFIDF (Term Frequency Inverse Document Frequency) method from Information Retrieval. It is called TFITF (Term Frequency Inverse Term Frequency) (Bonin, et al. 2010) and is calculated for terms extracted from one corpus and takes into account their frequencies in comparative data. Contrastive Selection via Heads (Basili, et al. 2001) is a method based on the distribution of the term head elements in corpora from several domains. 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.

TermoPL allows us to compare such a list with another list obtained using the same method from a different corpus and for common terms, the program indicates for which corpora they are more representative. But the results of this comparison, for not big corpora we worked with, were often not reliable. For example, some generally used expressions tend to be used more frequently in some types of texts. In our comparison of the medical ChH corpus with the general NKJP 1-million subcorpus, the LL method gave the same results for *dokumentacja medyczna* 'medical records' and *gruba warstwa* 'thick layer', the first one is a medical term and the second one is a general one. To make the decisions more reliable, we compare several (not necessary very big) corpora to gain the necessary information out of many comparisons. We analyse three different solutions to this problem and compare them on the same set of corpora.

Method I: co-occurrence in multiple corpora

The simplest approach for detecting irrelevant phrases could be identification of phrases which occur in more than one terminology list. To test this hypothesis, we check multi-word phrases which occur in more than three out of six tested corpora.

This approach has a drawback: We may be able to identify a very small number of phrases, if we decide to accept only those that occur in all but one corpora. For the less frequent phrases, we quickly get much less reliable candidates. The number of shared multi-word phrases is given in Table 2, while examples of such phrases are given in Table 3.

Table 3 Shared multi-word phrases, examples

Phrases common in 6 corpora

Irrelevant phrases	Domain Phrases
cecha charakterystyczna 'characteristic feature' *mały stopień* 'small degree'	*miejsce zamieszkania* 'place of residence'

Phrases common in 5 corpora

Irrelevant phrases	Domain Phrases

brak czasu ‘lack of time’	*historia sztuki* ‘history of art’
punkt zwrotny ‘turning point’	*gospodarka rynkowa* ‘market economy’
niski poziom ‘law level’	*chłop pańszczyźniany* ‘peasant serf’

Phrases common in 4 corpora

Irrelevant phrases	Domain Phrases
duża skala ‘large scale’	*akademia sztuk pięknych* ‘Academy of Fine Arts’
dokładna analiza ‘thorough analysis’	*tkanka tłuszczowa* ‘fat tissue’
elementarna potrzeba ‘elementary need’	*półkula mózgu* ‘hemisphere’
fizyczny kontakt ‘physical contact’	*grupa etniczna* ‘ethnic group’

The second issue is that we treat equally phrases that occur very frequently and those which are very rare. But if we set up a threshold on the number of occurrences, the number of shared phrases decreases dramatically in the case of small corpora. Table 2 shows that the number of phrases which occur at least twice is about seven times smaller than those occurring at least once.

Method 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. Since considered corpora have noticeably different sizes, we have to recalculate C-values to make them comparable in all data sets. For this purpose, we normalize C-values, as shown below, so that they sum up to some fixed number *N* in all corpora. We assume that *N* equals to 100,000. For a term *t* in corpora *C* we define:

$$C\text{-}value\text{-}norm(t) = \frac{N}{\sum_{t' \in C} C\text{-}value(t')} \times C\text{-}value(t)$$

We map the normalized C-values to the following five values (*C-map-value*):

- -1 if a term is not present in a corpus;
- 0.5 if a term has a normalized C-value near 0, in our experiment below 0,00001;
- 1 if the normalized C-value is below 1;
- 2 if the normalized C-value is below a selected threshold, equal to 8 in the experiment;
- 3 if the normalized C-value is above the threshold.

The threshold have been chosen on the basis of inspecting several term list in which there are a lot of very small values which we wanted to differentiate while bigger values were treated as indication of domain dependence. Then, we count the standard deviation (denoted by δ in the next two formulas) between mapped C-values of a term in all corpora and order terms according to the ascending values of the M_{II} coefficient, defined as follows:

$$M_{II}(t) = \frac{\sum_{all\ corpora} \delta_{(C-map-value(t))}}{number\ of\ corpora}$$

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 similarly useful in all corpora and their relative position from the top of the list is roughly the same.

In the modified version of the method, named IIa, we used log_{10} of the normalized C-values (*C-value-norm*) instead of the rigid five values (*C-map-value*) (still -1 was assigned to non-present terms). Table 4 gives examples of phrases, their C-values in each corpus and the values of the M_{II} and M_{IIa}coefficients.

$$M_{IIa}(t) = \frac{\sum_{all\ corpora} \delta_{log10(C-value-norm(t))}}{number\ of\ corpora}$$

Table 4 Examples of phrases, their C-values and $M_{II}, M_{IIa}(t)$ values

	ChH	HS	KS	Lit	Music	wikiE	#corpora	$M_{II}(t)$	$M_{IIa}(t)$
wysoki poziom `high level'	3.0	3.0	3.0	3.0	3.0	3.0	6	0.00	0.19
mały stopień 'small degree'	2.0	3.0	3.0	3.0	2.0	3.0	6	0.47	0.44
wysiłek intelektualny 'intellectual effort'	1.0	-1.0	1.0	1.0	1.0	0.5	5	0.73	0.5
hipoteza badawcza 'research hypothesis'	-1.0	0.5	0.5	1.0	-1.0	-1.0	3	0.85	0.6
historia sztuki 'history of art'	-1.0	3.0	3.0	3.0	1.0	2.0	5	1.46	1.23

Method III: penalization for not occurring in other corpora

Another method is based on the observation made in (Lopes, Fernandes and Vieira 2016), where it is suggested that terms appearing in the contrasting corpora should have been penalized proportionally to the number of their occurrences in considered corpora. Thus, the absolute frequency of the term in the domain corpus is divided by a penalization factor M_{III} given below and described in details in (Lopes, Fernandes and Vieira 2016). We adapted the idea proposed for selecting domain specific terms to calculate a list of irrelevant phrases ordered by a penalization factor based on term C-value instead of frequency as in the original paper. The higher the penalization factor, the lower the probability that the term is domain related. Table 5 gives results of the M_{III} coefficient for the same phrases as in Table 4.

$$M_{III}(t) = \prod_{\forall corpus\ C} (1 + log_{10}(C\text{-}value\text{-}norm^C(t))),$$

where *C-value-norm^C(t)* is the normalized C-value coefficient of term *t* calculated in corpus *C*.

Table 5 Examples of phrases with $M_{III}(t)$ values

		M_{III}(t)
wysoki poziom	'high level'	133.261
mały stopień	'small degree'	38.485
historia sztuki	'history of art'	22.721
wysiłek intelektualny	'intellectual effort'	2.042
hipoteza badawcza	'research hypothesis'	1.210

II+III, IIa+III second order methods

When analysing 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 (and IIa) and the III methods are the smallest. Thus, we combined weights obtained by these two methods in one by means of linear combination of their values normalized to the [0, 1] interval. 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. The results of the $M_{IIa+III}$ coefficient for the phrases considered in the previous two sections are given in Table 6.

$$M_{IIa+III}(t) = \alpha\big(1 - M_{IIa\text{-}norm}(t)\big) + (1-\alpha)M_{III\text{-}norm}(t)$$

Table 6 Examples of phrases with $M_{IIa+III}(t)$ values

		$M_{IIa+III}(t)$
wysoki poziom	'high level'	0.986
mały stopień	'small degree'	0.487
wysiłek intelektualny	'intellectual effort'	0.306
hipoteza badawcza	'research hypothesis'	0.273
historia sztuki	'history of art'	0.197

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. On the contrary, general terms and discourse expressions can accompany expressions from many unrelated domains, and hence they tend to have much more diverse contexts. To measure this diversity, we apply a clustering coefficient described in (Hamilton, Leskovec and Jurafsky 2016). It measures a word's contextual diversity, and thus polysemy.

5.1 Context diversity coefficient

In the method IV, we ordered all terms according to the increasing diversity coefficient. For each term in corpus T[3], the method creates the set of contexts. The context of a term consists of x words (in our experiment x=5) before and after the term. Then it measures the percentage of highly related pairs of elements in this set. A related pair of words is defined as a pair which has a non-zero Positive Pointwise Mutual Information (PPMI) value. The diversity coefficient is defined as follows:

$$M_{IV[T]}(w) = \frac{\sum_{c_i, c_j \in N_{PPMI}(w)} C_{N_{PPMI}}(c_i, c_j)}{|N_{PPMI}(w)|(|N_{PPMI}(w)| - 1)},$$

where $C_w = \{w_i : w_i \; is \; in \; a \; context \; of \; w \; in \; corpus \; T\}$,
$N_{PPMI}(w) = \{w_j \in C_w : PPMI(w, w_j) > 0\}$,
$C_{N\_PPMI}(c_i, c_j) = \{1 \; if \; PPMI(c_i, c_j) > 0, \; and \; 0 \; otherwise\}$.

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 word pairs, the stronger correlation.

$$PPMI(w, z) = \max\left\{log \frac{p(w, z)}{p(w) * p(z)}; 0\right\}$$

The tested hypothesis was whether the lower coefficient of the method IV indicates more domain related terms which are less polysemous. As in principle, a irrelevant phrases could not have any highly related contexts, we suggest modifying the above coefficient by replacing the nominator by the number of all possible context pairs. The modified coefficient is defined as follows:

$$M_{IV'[T]}(w) = \frac{\sum_{c_i, c_j \in N_{PPMI(w)}} C_{N\_PPMI}(c_i, c_j)}{|C_w|(|C_w| - 1)}$$

Another modification of the $M_{IV[T]}(w)$ coefficient concerns the weight assigned to the context words. Instead of treating all of them uniformly, we assigned the highest weight 1.0 of the nearest neighbours of the word, and gradually decreased this weight for more distant words, For

3 Corpora used by method IV are given in square brackets.

example, in our experiment, we decreased the weight by twenty percent, so the last (fifth word) to the left and to the right have the weight equal to 0.2. We indicate this variant adding description *wght* to the method name.

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 the method IV is introduced. For such a case, we propose an additional step for selecting terms which are similar to the analysed one. Similarity is defined here as the cosine similarity of the vectors from the word2vec model (Mikolov, Yih and Zweig 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 (a value chosen experimentally). We observed that, for multi-word terms, the similarity coefficient is generally lower than for one-word terms and that, in a small corpus, the higher threshold provides very few similar terms. In Tables *Table* 7-9, 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 9 rather contains terms unrelated to the considered one, i.e. *dzieło stworzenia* 'act of creation'.

Table 7 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 8 Similar multi-word terms for *dziewiętnasty 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'

Table 9 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'

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

$$M_{IVadd[T]}(w) = M_{IV'[T]}(\{w' : sim(w, w') > 0.44\})$$

To conclude, we defined two basic variants of the method counting diverse contexts of terms: M_{IV} and $M_{IV'}$. The first one is based on the clustering coefficient described in (Hamilton, Leskovec and Jurafsky 2016), the second is a modification of this coefficient consisting in taking into account all possible context pairs instead of those which have a non-zero PPMI. Both above methods can have a variant with added *wght* description which treats differently closer and further contexts and a variant with *add* description which takes into account phrases similar to the considered term. Finally, for all these methods, contexts can be counted on various corpora, the name of a corpus is given in square brackets.

5.2 Boosting lists of irrelevant phrases by adding similar ones

The last method of identification of irrelevant phrases uses distributional models more directly. The list of the most similar phrases cited in Tables 7-8 shows that phrases similar to an irrelevant one tend to be also irrelevant. As the previously described methods produce ranked lists of allegedly irrelevant phrases, we can assume that the top part of the list contains a larger proportion of irrelevant phrases than the lower parts. We can use them as seeds for collecting other irrelevant phrases from the most similar terms. Similarity is calculated as cosine similarity of vectors from the distributional semantic model trained on the bigger corpus used to evaluate the method IV. In this method, we obtain binary information if a given phrase is irrelevant or not. We are not able to rank new candidates but we can influence the results choosing various thresholds based on similarity or term ranking.

According to the method V, a phrase is irrelevant (M_V^X=1) if it is selected by the method X as irrelevant ($rank_X(w) < K$) or is similar to such an element of this list. The method X can be any of the previous methods.

$$M_V^X(w) = \begin{cases} 1 & if \ \ (rank_X(w) < K) \ or \ \ (\exists\, w_i \ rank_X(w_i) < K \ \ and \ \ \ sim[T](w, w_i) < \theta\ ), \\ 0 & otherwise \end{cases}$$

where X is the method identifying irrelevant phrases;
K is a threshold used to indicate irrelevant phrases on the list generated by X;
θ is a threshold used to select similar phrases by a distributional semantic model trained on the corpus T.

6. Evaluation

6.1 Evaluation data

To evaluate our methods, we prepared two manually annotated lists. The first one, called *COM*, consists of 7001 terms which occur in at least three of the six selected corpora. Annotation was done by two annotators (computer scientists working for several years on computational terminology extraction) and then the third one resolved the conflicts. The annotators introduced five labels representing non-terms, general terms, domain terms used generally, domain terms,

and improper phrases. Finally, if an annotator is not able to make a decision s/he can assign '?'. The annotators received the following 10 rules, which should be checked in the given order. If a given rule is satisfied, the appropriate annotation is assigned without checking the remaining points.

1. Geographical names and names of people (named entities) are general terms. The interpretation of phrases containing named entities depends on co-occurring words, e.g. *nagroda Nobla* 'Nobel Prize' is the general term, but 'pieśń Schuberta' 'Schubert song' is the domain term.
2. If a phrase explicitly indicates one (or at most two) of our domains, then it is considered to be a domain term, e.g. *akumulacja kapitału* 'accumulation of capital', *rzeźba antyczna* 'antique sculpture' *tragedia antyczna* 'antique tragedy', *chłop pańszczyźniany* 'peasant serf'.
3. If a phrase clearly points to a field out of the considered domains, then it is considered as a domain term used generally, e.g. aparat fotograficzny 'camera' or *dach domu* 'home roof'.
4. If a term is wrong e.g. 'Anda the'[4] or truncated *czas Ludwika* 'time of Louis', *artykuł opublikowany* 'article published in', *chłopiec urodzony…* 'boy born …' it is annotated as an improper phrase.
5. Discourse markers are non-terms, e.g. *punkt widzenia* 'point of view', *kluczowe zagadnienie* 'key issue'.
6. Metaphors are non-terms, e.g. *ciężkie serce* 'heavy heart', *chleb powszedni* 'daily bread'.
7. Phrases which cannot be considered as a term (their meaning depends on the context) are non-terms, e.g. *jakieś słowo* 'any word', *poprzednia rata kredytu* 'previous loan instalment'.

[4] It is the part of improperly tagged English phrases used in Polish texts and containing 'and the'.

8. Abstract phrases commonly used in different texts are considered general terms, e.g. *ogląd rzeczywistości* ‘view of reality’.
9. If a phrase can be used in many domains, it is considered as a general term e.g. *aktywny udział* ‘active participation’, *wnikliwa analiza* ‘careful analysis’, *charakterystyczna cecha* ‘characteristic feature’.
10. If an annotator doubts whether a phrase is a domain or non-domain term, s/he should label it as a general term unless s/he is totally not able to make a decision and should then assign ‘?’.

Despite the guidelines cited above, the task turned out to be difficult and quite often the annotators disagree in their phrase ratings. The analysis of the data shows that the assumption that multi-word phrases are most often unambiguous and ambiguity refers to the negligible number of phrases is correct only partially as many two-word phrases have more than one meaning. Most of them are phrases with both a literal and metaphorical meaning, e.g. *prawa ręka* ‘right hand’ in medical texts has the literal meaning; in texts about politics, it has a metaphoric meaning as it refers to an important assistance; in sport texts, it occurs in both meanings. A few phrases have more than one literal meaning, for example *dalszy plan* means ‘long-range plan’ in many texts, but in the art domain, it means ‘background’. The first meaning refers to a general term, while the second one to a domain term. Another reason for disagreement among annotators derives from popular phrases used so often in everyday language that they can be rated as having a general meaning, such as *wielka polityka* ‘great politics’ and *dziedzina nauki* ‘scientific domain’. As some of the analysed corpora contain a lot of journalistic papers, such phrases are often found in our data. Finally, the annotators differently rate phrases that might be considered as truncated, such as *członek rady* ‘member of the board’, which probably should contain more information about the type of the ‘board’, or *próba dojścia*, which is annotated as an error by one annotator (probably as a part of a longer phrase ‘attempt to reach something’ and as a general one by the second annotator (probably as a

literal phrase ‘attempt to arrive’). (The very similar phrase *próba powrotu* ‘attempt to return’ is rated by both annotators as the general one.)

All these issues are reflected in a relatively low Cohen's kappa-coefficient which was equal to 0.4. At the evaluation stage, we treated the first three classes (i.e. non-terms, general terms and domain terms used generally) together as irrelevant phrases, which did not change the Cohen kappa-coefficient very much (increased to 0.45). Table 10 includes the number of annotations of each type. As the problem looks difficult, we decided to check the stability of phrase ratings by the same annotators after several months. They verified the original annotations. The final version differs from the original one on about 10% labels. As *COM* test set contained a lot of phrases located very low on the ranked terminological lists, we also prepared the second test set 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.[5] 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 (*MFQ*) contains only 2341 terms which were annotated identically by both annotators. 730 terms are included in both *COM* and *MFQ* sets.

Table 10 Manual annotation

	COM test set			*MFQ* test set annotations		
	An1	An2	agreed	An1	An2	common *MFQ*
general terms	6128	5141	4839	1493	1296	999
non-general terms	736	1572	1910	1571	1893	1342
improper phrases	115	266	252	175	51	–

6.2 Results

As our results are ranked lists, we had to introduce a threshold indicating which part of the lists should be treated as irrelevant phrases. For the first method, we selected terms which occur in at least 4 corpora; for the others, we treated the top 70% of the lists as irrelevant phrases. This is roughly the most desirable partition, as the annotation of *COM* test set contains about 69% of

[5] Most of the top terms for this corpus are domain specific, see (Marciniak and Mykowiecka, 2015).

irrelevant phrases (72% if we also count errors). We compared the annotations done by each method with the *COM* standard, moreover we compared the annotations for each pair of the methods. The results are given inTable 11

Komentarz [A1]: Czy ktoś pamięta czy przy porównywaniu metod braliśmy pod uwagę wartości w COM (GS) czy po prostu wszystkie wspólne?

Table 11 Common annotations for *COM* test set done by annotators for each method separately and for pairs of the methods

method	I	II	IIa	III	IIa+III	IV[art]	IV[nkjp+art]
GS	2827	4520	4480	3996	6410	3713	3909
I	-	2643	2623	3637	2603	3269	3132
II	-	-	4539	2799	4721	3586	3192
IIa	-	-	-	2799	4701	3559	3185
III	-	-	-	-	3978	4353	4720
IIa+III	-	-	-	-	-	3622	3206
IV_{art}	-	-	-	-	-	-	3385

For the evaluation of the IV method, we performed experiments in which we used two data sets. of different sizes and specificity. The first (*art*) corpus consisted of four of the corpora described in Section 3 (except the hospital data set – ChH and the economy corpus – wikiE). It consists of about 845K tokens. Thus it is a small, specialized corpus. The second data set (*nkjp+art*) is a general one and it 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 test term list is the same list of 7001 terms described above as the *COM* set. While counting the diversity coefficient(method IV), we selected contexts containing only lower case letters; thus, we excluded named entities from this set. We also disregarded contexts which are 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.

For all methods, we counted how many terms annotated as irrelevant in the *COM* file were found in the consecutive parts 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.

Figure 1 Percentage of irrelevant phrases for every 500 terms individually for the methods I-III , *COM* test set. The numbers depicted on the figure show how many of these phrases were

located on the positions 1-499, 500-999 and so on, on the ordered lists obtained by each methods.

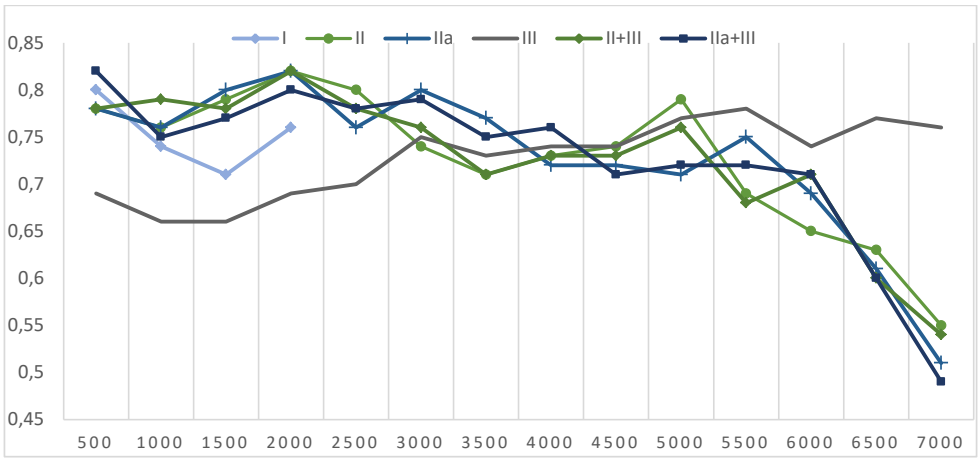
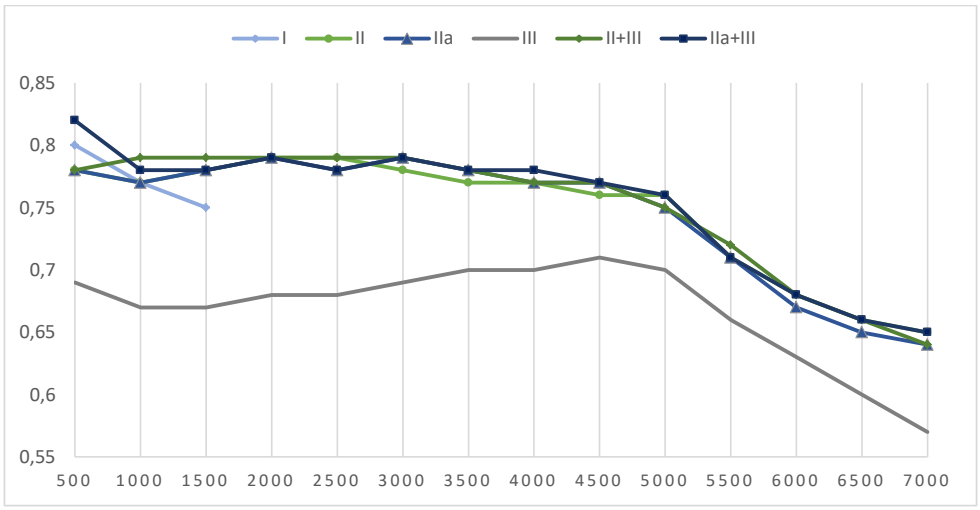


Figure 2 Cumulative precision of the methods I-III, *COM* test set



Figures 1 and 2 show that the methods II, IIa, II+III, IIa+III do not differ much. The most stable results were achieved for the combination IIa+III. For this method, we tested several values of α from 0.2 to 0.8 and the best results were obtained with α=0.4. In the results of method III, irrelevant phrases are nearly equally spread with even the increasing tendency. The precision of this method is hence lower than others.

Next two figures (3 and 4) show the comparison of the best method based on domain corpora (IIa+III) with the method based on term contexts (IV). The top parts of the lists obtained by these methods contain more irrelevant phrases than the list produced by the IIa+III method.

Figure 3 Percentage of irrelevant phrases for every 500 terms individually for different variants of the method IV compared to IIa+III – *COM* test set

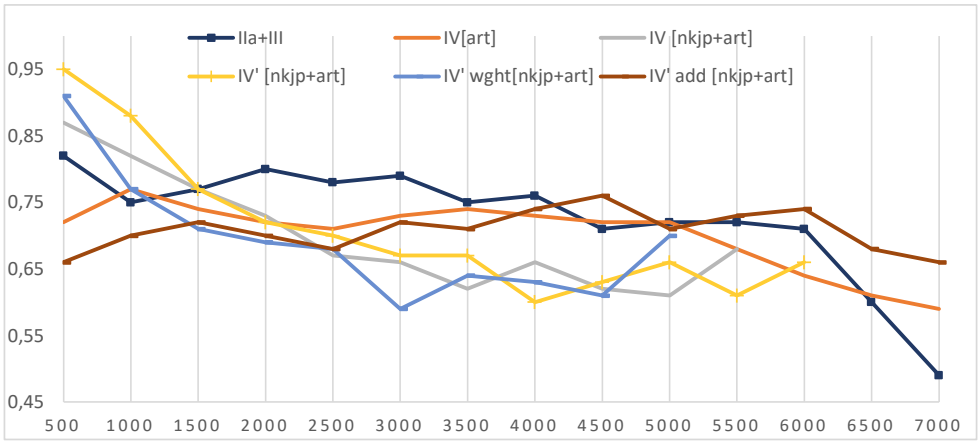
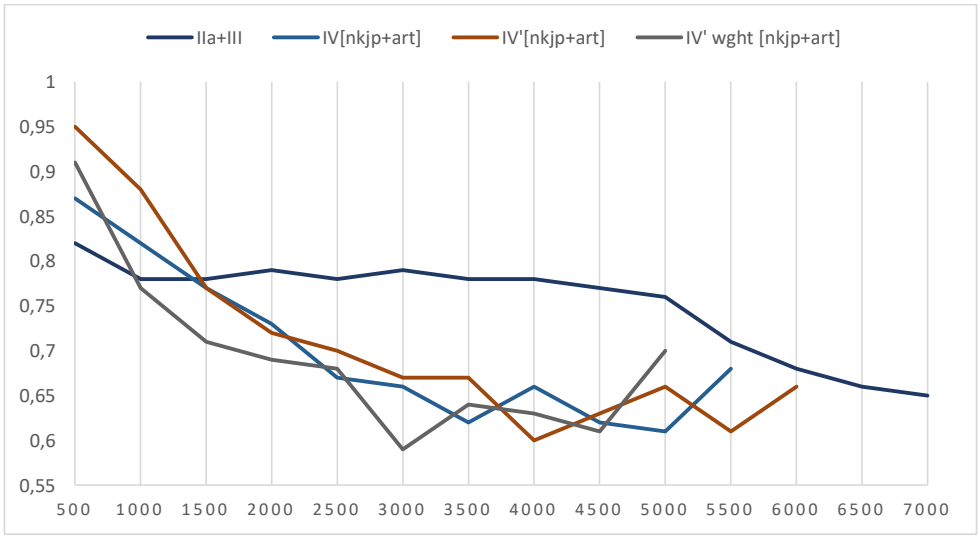


Figure 4 Cumulative precision of methods for different variants of the method IV compared to IIa+III – *COM* test set



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. For this data set, the addition of similar terms (IVadd[art]) 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 0.69 for the entire set. For the big corpus, the results achieved by adding similar terms (IVadd[nkjp+art]) were slightly worse, as was expected. Table 12 summarizes the results and shows 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).

Table 12 Precision of all the methods – *COM* test set

	I	II	IIa	III	II+III	IV [art]	IVadd [art]	IV [nkjp+art]	IVadd [nkjp+art]
first 500 terms	0.80	0.78	0.78	0.69	0.78	0.56	0.64	0.87	0.67
entire list	0.41	0.65	0.64	0.57	0.64	0.56*	0.69*	0.68*	0.60*

To check whether the methods behave in the same way on different test sets, we used them on the subset of the *MFQ* set described before (the sum of 1000 most frequent phrases in our corpora) containing phrases which occurred in at least two of our data sets (*MFQ-min2*). There are 1187 such phrases. The phrases in this set are rather domain terms as they appear on the top of the appropriate term candidates lists, but as they occur in more than one corpus they might also be out-of-domain terms. In fact, only 445 phrases on this list are domain terms. We expect our methods to move out-of-domain phrases to the top of the produced ranked list. Table 13 depict the percentage of such terms in every 250 elements' part of these lists. We can observe that methods I and IV are most consistent in this respect for all data sets.

Table 13 Percentage of out-of-domain terms in the subsequent parts of the result lists obtained by different methods.

	COM test set					*MFQ-min2* test set				
positions	I	II	III	IIa+ III	IV $_{nkjp+arr}$	I	II	III	IIa+ III	IV $_{nkjp+art}$
1- 250	0.81	0.80	0.72	0.86	0.90	0.86	0.81	0.83	0.79	0.83
250- 499	0.80	0.77	0.66	78	0.84	0.72	0.52	0.68	0.68	0.68
500- 749	0.77	0.72	0.68	0.78	0.84	0.45	0.47	0.59	0.65	0.50
				0.76						

750- 999	0.72	0.81	0.64	0.76	0.79	0.50	0.67	0.44	0.59	0.41
1000-1249	0.70	0.80	0.69	0.74	0.78	0.58	0.67	0.56	0.35	0.44
1250-1499	0.72	0.77	0.64	0.80	0.77	-	-	-	-	-

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 from the basic version of the method, we tested the method IV' and the non-uniform treatment of the context words (IV'wght method). We performed tests on the big NKJP together with ART corpus. The results shown in **Błąd! Nie można odnaleźć źródła odwołania.** confirm improvement in cases where the method IV'wght was used. Figure 4 shows the cumulative precision at each 500 terms counted for different variants of the method IV. The number of irrelevant phrases 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 the results to deteriorate. In comparison to the methods II-III, the method IV showed the quickest decrease of the percentage of the irrelevant phrases for each five hundred positions, thus proving to be the most selective one.

Table 14 Percentage of irrelevant phrases in the subsequent parts of the result lists obtained by different variants of the method IV used on the nkjp+art corpus

	COM test set			*MFQ-min2* test set		
positions	IV	IV'	IV' wght	IV	IV'	IV' wght
1- 250	0.90	0.96	0.94	0.83	0.97	0.92
250- 499	0.84	0.95	0.89	0.68	0.70	0.65
500- 749	0.84	0.90	0.78	0.50	0.55	0.53
750- 999	0.79	0.88	0.76	0.41	0.36	0.31
1000-1249	0.78	0.60	0.74	0.44	0.28	-

Table 15 shows how many irrelevant phrases were filtered out from the top part of terms in the 5 domain corpora. We tested lists of 1800 top irrelevant phrases obtained by selected 8 methods separately. We tested only the top parts of all domain term lists consisting of 10K terms. Table 16 shows how many false positive irrelevant phrases are filtered out under the same conditions. In these two tables we can observe that the method I proved to be quite effective, but it is of

limited practical use as it can address only the small subset of terms. It does not change the ordering of the entire candidates list. The 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 phrases: irrelevant and false positive irrelevant ones. The best results are obtained by the method IV'.

Table 15 Irrelevant phrases filtered out of 10K top terms

corpus	I	II	IIa	III	IIa+III	IV[art]	IV[nkjp+art]	IV'[nkjp+art]
ChH	61	30	30	**65**	39	1	39	43
HS	344	23	46	**451**	128	23	216	290
Music	371	26	71	**446**	148	46	285	402
Lit	606	31	76	**771**	180	33	491	621
wikiE	255	24	65	**294**	143	16	169	219
KS	533	29	69	**771**	167	27	426	527

Table 16 Filtered false irrelevant phrases in 10K top terms

corpus	I	II	IIa	III	IIa+III	IV[art]	IV[nkjp+art]	IV'[nkjp+art]
ChH	24	5	3	**39**	8	0	6	6
HS	119	1	6	**209**	29	21	69	53
Music	98	0	13	**132**	41	10	47	52
Lit	211	4	13	**353**	48	26	124	115
wikiE	106	3	14	**179**	35	19	86	96
KS	218	2	17	**369**	52	38	172	173

In the last experiment we evaluated the method described in section 5.2 in which we propose to boost a list of irrelevant phrases by similar phrases. To recognize phrases similar to a considered one, we tested four distributional models generated from the data set containing the NKJP and ART corpora. Word vectors were trained using gensim implementation of word2vec (Řehůřek and Sojka 2010). As seeds we decided to use a subset of *COM* and *MFQ* sets. To obtain vectors for multi-word phrases we chose two techniques. The first one directly combines vectors of single words constituting multi-word phrases from *COM* and *MFQ* into their sum (first model, *sum*) or dot product (second model, *mult*). The other technique substitutes all terms from *COM* and *MFQ* by their unique identifiers in every sentence from input data set before the process of training. In every sentence as many substitutions takes place as possible. This might yield the

situations where a few sentences are produced out of one as multi-word phrases may overlap, e.g. *high artistic level* and *artistic level*. For training we chose only sentences where at least one substitution had taken place (third model, *mwt*) and the above-mentioned substituted sentences together with all sentences from NKJP and ART (fourth model, *all sentences*).

We selected two lists with high probability of containing irrelevant phrases. The first one is the list of 357 phrases which occurred in at least five of the tested corpora (*6-5-list*). The second one consists of nearly the same number of top phrases from the list obtained by the II+III method *(358top-list*). For every phrase contained in these two lists, we selected top five most similar phrases with the cosine similarity of at least 0.4. The results are depicted in Table 16 in which the percentage of the irrelevant phrases which are present in the *COM* manually annotated test set together with the percentage of domain terms, the total number of identified phrases, the number of one-word terms identified, and the number of multi-word phrases which are not present in *COM* test set are given. The results do not differ much, but this observation confirms the *mult* model is more selective for multi word terms, thus in practice gives more good candidates than the others. To see how the better quality of the initial list influence the results we prepared sublists with irrelevant phrases: *6-5-non-domain* and *358top-non-domain*. As it was assumed, the percentage of the domain terms in the results was significantly lower while the absolute number of new irrelevant phrases did not decrease much. This observation leads to the conclusion that this method of irrelevant phrases set population should be used for already manually checked data. This approach can solve more effort needed to verify longer lists with more incorrect elements.

Komentarz [A2]: W COM czy w COM + MFQ jak piszemy wcześniej, że stamtąd bierzemy dane?

Komentarz [A3]: Mam nadzieję że z obu

Table 17 The results of boosting lists of terms by adding similar phrases using different vector models.

model	Percentage of out-of-domain terms	Percentage of domain terms	Number of multi-word phrases	Number of one-word terms	Number of phrases out of the scope of *COM* list (not evaluated)
6-5-list					
1. sum	0.723	0.198	1293	239	101
2. mult	0.746	0.197	1257	0	89
3. mwt	0.750	0.189	1375	153	83
4. all sentences	0.746	0.196	1372	134	79
6-5-list-non-domain					
1. mult	0,880	0.061	998	0	58
2. all sentences	0.888	0.063	1082	93	52
358top-list					
3. sum	0.647	0.228	1329	452	166
4. mult	0.663	0.229	1207	2	130
5. mwt	0.685	0.219	1336	447	127
6. all sentences	0.681	0.222	1366	409	133
358top-non-domain					
1. mult	0.826	0.104	917	2	64
2. all sentences	0.842	0.101	1029	249	58

7. Conclusions

Differentiation between irrelevant phrases and domain specific terms is a hard task. It is difficult not only for computers but for human annotators too, as the Cohen's kappa-coefficient between annotators is low. It seems reasonable to consider the method of preparation of the gold standard involving more annotators and aggregation of the results (following e.g. SimLex-999 data annotation (Hill, Roi and Korhonnen 2015)). Moreover, it may be useful to design methods dedicated to recognition of different types of irrelevant phrases separately. The methods proposed in this paper allow for preselecting sets of phrases containing more than seventy percent of irrelevant phrases.

For the methods based on domain corpora, the most efficient and, at the same time, simple method relies on standard deviation for the C-value coefficient (method IIa). Unfortunately, these methods recognize many infrequent terms. The method III operates on the top part of automatically extracted term lists, but its precision is lower than the other methods.

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 larger 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 $M_{IV'}$ 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 which is consistent with our expectations.

The methods described in our paper can be used to select candidates for irrelevant phrases. Such a set can help when preparing lists of concepts shared by several domains. However, its usage for the task of eliminating unwanted terms from the terminological list obtained automatically is limited, as the precision of the method is not very high so manual verification of the list is recommended. Expansion of the list by adding similar phrases could be a good method for identifying irrelevant phrases which are similar to commonly used ones.

8. Bibligraphy

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

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

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima. “Automatic recognition of multi-word terms: the C-value/NC-value Method.” *Int. Journal on Digital Libraries*, 2000: 115-130.

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

Hill, Felix, Reichart Roi, and Anna Korhonnen. “Simlex-999: Evaluating semantic models with (genuine) similarity estimation.” *Computational Linguistcs*, 2015.

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

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

Marciniak, Małgorzata, and Agnieszka Mykowiecka. “Terminology extraction from medical texts in Polish.” *Journal of Biomedical Semantics*, 2015.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. “Linguistic Regularities in Continuous Space Word Representations.” *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings.* Atlanta, Georgia: The Association for Computer Linguistics, 2013. 746-751.

Navigli, Roberto, and Paola Velardi. “Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites.” *Computational Linguistics*, 2004: 151-179.

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

Rayson, Paul, and Roger Garside. “Comparing Corpora Using Frequency Profiling.” *Proceedings of the Workshop on Comparing Corpora - Volume 9, WCC ’00.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 1-6.

Řehůřek, Radim, and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora.” *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valetta, Malta: ELRA, 2010. 45-50.

Schäfer, Johannes, Ina Rösinger, Ulrich Heid, and Michael Dorna. “Evaluating Noise Reduction Strategies for Terminology Extraction.” *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence.* Granada: Universidad de Granada, 2015. 123-131.