NPMI driven recognition of nested terms

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

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

Abstract

In the paper, we propose a new method of identifying terms nested within candidates for the terms extracted from domain texts. The list of all terms is then ranked by the process of automatic term recognition. Our method of identifying nested terms is based on two aspects: grammatical correctness and normalised pointwise mutual information (NPMI) counted for all bigrams on the basis of a corpus. NPMI is typically used for recognition of strong word connections but in our solution we use it to recognise the weakest points within phrases to suggest the best place for division of a phrase into two parts. By creating only two nested phrases in each step we introduce a binary hierarchical term structure. In the paper, we test the impact of the proposed nested terms recognition method applied together with the C-value ranking method to the automatic term recognition task.

1 Introduction

The Automatic Term Recognition (ATR) task consists in identifying linguistic expressions that refer to domain concepts. This is usually realised in two steps. In the first one, candidates for terms are identified in a corpus of domain texts. This step usually consists in identifying grammatically correct phrases by means of linguistically motivated grammars describing noun phrases in a given language. However, sometimes no linguistic knowledge is utilised and candidates for terms are just frequent n-grams as in (Wermter and Hahn, 2005). The second processing step consists in ranking the extracted candidates and selecting those which are most important for a considered domain. This task is usually based on statistics.

The ranking procedure can be based on different measures which are characterised as either "termhood-based" or "unithood-based". Kageura and Umino (1996) defined the termhood-based methods measure as "the degree that a linguistic unit is related to domain-specific concepts", i.e. the likelihood that a phrase is a valid domain term. The unithood-based methods measure the collocation strength of word sequences, usually with the help of log-likelihood, pointwise mutual information or T-score measures, described in (Manning and Schütze, 1999), while ATR applications based on them are described in e.g., (Pantel and Lin, 2001), (Sclano and Velardi, 2007). A comparison of these approaches is given in (Pazienza et al., 2005). Some hybrid solutions to the ATR problem have also been proposed (Vu et al., 2008) or (Ventura et al., 2014). In the paper (Korkontzelos et al., 2008), the comparison between these two groups of methods led the authors to the conclusion that the termhood-based methods outperform the unithood-based ones.

This paper is devoted to the problem of selecting candidates for terms from an annotated domain corpus. Our approach is based on the C-value method, (Frantzi et al., 2000). An important feature of this method that attracted our attention was the focus on nested terms. Frantzi *et al.* (2000) described nested terms as terms that appear within other longer terms, and may or may not appear by themselves in the corpus. They show that recognition of nested terms is very important in terms extraction, but they also give examples when a nested phrase constructed according to the grammar rules is not a term. One of

Proceedings of the 4th International Workshop on Computational Terminology, pages 33–41,
Dublin, Ireland, August 23 2014.

these examples is the phrase *real time clock* which has two nested phrases: *real time* and *time clock*, but the second one is not a good term. The authors define the C-value measure that is used to rank candidate terms extracted from a domain corpus, together with their nested terms. It is counted on the basis of the frequency of the term as a whole phrase in the corpus, its frequency as a nested phrase in other terms, the number of different phrases in which that nested phrase occurred, and its length. The authors expect that phrases that aren't considered as terms should be placed at the end of the list ordered according to this coefficient value.

We applied the C-value method to extract terminology from a corpus of hospital discharge documents in Polish. Experiments, where different methods of counting the C-value were tested, are described in (Marciniak and Mykowiecka, 2014). Unfortunately, a few grammatically correct but semantically odd phrases were always placed in the top part of the ranking list of terms. Examples of such phrases, placed among the 200 top positions, are: *USG jamy* 'USG of cavity' being a nested phrase of the very frequent phrase *USG jamy brzusznej* 'USG of abdominal cavity', *infekcja górnych dróg* 'infection of upper tract' or *powiększony węzeł* 'enlarged node'.

We propose a method that prevents the creation and promotion of such nested phrases to be considered as terms. The main idea is to use a unithood-based method e.g., Normalised Pointwise Mutual Information (NPMI) (Bouma, 2009) for driving recognition of nested phrases. Our solution is based on the division of each considered phrase into only two parts. The places where a phrase is divided must create nested phrases that are consistent with grammar rules. Usually, there are several possible places for division of a phrase. From all of them, we choose the weakest point according to NPMI counted for bigrams on the basis of the whole corpus. So, as a bigram constitutes a strong collocation, it prevents the phrase from being dividing in this place, and does not usually lead to the creation of semantically odd nested phrases, of which examples are given above.

The analysed corpus of Polish medical texts is described in Section 2. In the following two sections we present the method in detail. Then, in Section 5, we describe the comparison of the resulting lists of terms ranked according to the C-value measure, for two methods of recognition of nested phrases , i.e.: for all possible phrases fulfilling grammatical rules, and for the method proposed in the paper.

2 Corpus description

The domain corpus consists of 3116 hospital discharge documents gathered at a hospital in Poland. Texts came from six departments and were written by several physicians of different specialties. The collected texts were analysed using standard general purpose NLP tools. The morphological tagger Pantera (Acedański, 2010), cooperating with the Morfeusz analayser (Woliński, 2006), was used to divide the text into tokens and annotate them with morhosyntactic tags. They included a part of speech name (POS), a base form, as well as case, gender and number information, where they were appropriate. This information is used by shallow grammars recognising the boundaries of nominal phrases — term candidates and, also,sources for nested phrases. The corpus consists of about 2 million tokens in which a shallow grammar recognised more than 22 thousand noun phrases.

The corpus contains quite a lot of words unrecognised by Morfeusz as the vocabulary of the clinical documents significantly differs from general Polish texts. Additionally, the texts are not very well edited despite the spelling correction tools being usually turned on, so they contain quite a lot of misspelled words. This results in 22,000 unrecognised tokens (many of them are medications, acronyms and units) that are not taken into account when nominal phrases are recognised. Consequently, it lowers the number of phrases, and affects the quality of some of them. In (Marciniak and Mykowiecka, 2011), the problems of morphological annotation of hospital documents in Polish are presented and the reasons for the many unrecognised tokens are highlighted.

3 Nested phrases recognition

In this section, we describe the way to create a list of term candidates that takes into account nested phrases. This task is usually supported by linguistic knowledge that allows for identifying candidates for terms which are syntactically valid.

In the extraction step, we identified complex noun phrases consisting of nouns with adjectival and nominal modifiers obeying Polish grammar rules (in particular, case, gender and number agreement). The types of Noun Phrases under consideration can be schematically defined as below:

AdjPhrase Noun AdjPhrase
AdjPhrase Noun
Noun AdjPhrase
NounPhrase NounPhrase-in-genitive

Noun Phrases were extracted from the corpus using a cascade of shallow grammars. As Polish is a highly inflected language, we operate on simplified base forms of phrases in our computations, consisting of lemmas of subsequent words. This approach, proposed for ATR in Polish in (Marciniak and Mykowiecka, 2013), allows us to unify forms of phrases in different cases and numbers. For example: *przewlekłe zapalenie gardła*, *przewlekłe zapalenia gardła*, *przewlekłego zapalenia gardła*, *przewlekłych zapaleń gardła* are forms of 'chronic pharyngitis' in nominative singular and plural and genitive in both numbers.[1] The extracted phrases constitute a foundation for creating the list of term candidates. Then we add nested phrases, recognised within those phrases, to the list of term candidates. The rules for identifying nested terms are described in the rest of this section.

3.1 Motivations

The original C-value method (Frantzi et al., 2000) recommends that all grammatical phrases, created from the maximal phrases identified in a corpus, should be considered as term candidates. But using this method, we quite frequently obtain nested grammatical subphrases which are syntactically correct, but semantically odd. One such phrase is *infekcja górnych dróg* 'infection (of the) upper tract' that is created from *infekcja górnych dróg oddechowych* 'infection (of the) upper respiratory tract'.[2] The last phrase has many different longer phrases in which it is nested, eg: *(częsta, drobna, ostra, bakteryjna...) infekcja górnych dróg oddechowych* '(often, minor, acute, bacterial...) infection (of the) upper respiratory tract', but it always concerns *drogi oddechowe* 'respiratory tract'. We observe that the bigram *drogi oddechowe* 'respiratory tract' constitutes a strong collocation. So the original phrase shouldn't be divided in this place to create a phrase containing the word *drogi* 'tract' without adding its type, i.e., *oddechowe* 'respiratory' in this case. Nominal phrases are usually constructed from two parts (except for coordinated phrases and nouns with more complex subcategoriaztion frames, which usually do not fulfill agreement constraints in Polish). For nominal phrases from domain corpora, we suggest that the best place for the division is indicated by the weakest bigram.

After considering patterns of nominal phrases in Polish, we realised that the weakest connections are usually between two nominal phrases (the last pattern). So, an adjective more likely modifies the nearest noun and not the whole phrase, as in: $prawidłowa_{adj}$ $mikroflora_{noun}$ $górnych_{adj}$ $dróg_{noun}$ $oddechowych_{adj}$ 'normal microflora (of the) upper respiratory tract'. In this phrase, all the outermost adjectives are important parts of nominal phrases constructed around their nearest nouns, and it should be divided into two nominal phrases: *prawidłowa mikroflora* 'normal microflora' and *górne drogi oddechowe* 'upper respiratory tract'. However, it is not the universal rule. Let us consider another example: *częste infekcje górnych dróg oddechowych* 'frequent infections (of the) upper respiratory tract', where *częste* 'frequent' modifies the whole phrase. To account for this observation, we may slightly prefer divisions into two nominal phrases instead of an adjective and a nominal phrase.

3.2 Algorithm

From several methods for counting the strength of bigrams we chose the normalised pointwise mutual information proposed by Bouma, (2009), as it is less sensitive to occurrence frequency. We were looking for a method for which the bigram, consisting of a rare and a frequent token, will be high if the rare token only appears in connection with the frequent token, as, for example, for *esowate skrzywienie* 'S-shaped curvature'. The definition of this measure for the 'x y' bigram, where x and y are lemmas of sequence

[1] Further in the paper we will use phrases in the nominal case and singular number forms. These forms may differ slightly from the same phrases being nested ones (in genitive).

[2] The word order of the translation is different.

tokens, is given in (1), where p(x,y) is a probability of the 'x y' bigram in the considered corpus, and p(x), p(y) are probabilities of 'x' and 'y' unigrams respectively.

$$NPMI(x,y) = \left(ln \ \frac{p(x,y)}{p(x)p(y)} \right) \Big/ - ln \ p(x,y) \tag{1}$$

First, we extract all the grammatical phrases from the corpus, taking into account only the maximal one. Then, for each phrase we identify all places where it can be divided according to grammar rules. We count NPMI for those and indicate the weakest connection in the phrase. Then, we divide it into two parts in this position. There are two possible situations: the first, when the phrase is divided into two nominal phrases; the second, when one phrase is a nominal phrase while the second one is an adjective phrase. In the first case, we add both parts to the list of term candidates and process the obtained parts of the phrase in the same way. In the second case, only a nominal phrase is added to the list and only this phrase is further divided.

```
nested_phrases (phr)
  if length(phr) > 1
    find all i positions where phr can be divided according to the grammatic rules
    for all i positions
      count NPMI(i-th bigram of phr)
    sort NPMIs from the lowest to the highest value
    j := position with the lowest NPMI
    if the j-th position divides phr into two nominal phrases
      divide phr into phr1 and phr2 on j-th position
      add phr1 and phr2 to the list of nested terms
      nested_phrases(phr1)
      nested_phrases(phr2)
    else
      n := position with the lowest NPMI where phr is divided into two nominal phrases
      if (120% NPMI(j)) > NPMI (n)
        divide phr into phr1 and phr2 on n-th position
        add phr1 and phr2 to the list of nested terms
        nested_phrases(phr1)
        nested_phrases(phr2)
      else
        if phr is divided on j position into adjective phrase to the left of nominal phrase
          cut off the outermost left element from phr
        else
          cut off the outermost right element from phr
        add phr to the list of nested terms
        nested_phrases(phr)
```

Figure 1: Procedure of nested phrases recognition

To take into account the specificity of adjectives in Polish nominal phrases described in 3.1, we decided to introduce a slight modification to the basic algorithm. If the weakest connection prefers the cutting of an adjective part from a phrase, we find the nearest place where the phrase is divided into two nominal phrases. Then, we compare the NPMI value referring to this bigram with 120% (fixed experimentally) of the lowest NPMI value. If it is still lower, we cut off one outermost element (adjective or adverb) from this adjectival part of the phrase and add the slightly shorter phrase to the term list. In other case, we divide the original phrase in that second place into two nominal phrases. The algorithm is given in Figure 1.

The grammatically correct nested phrases				The nested phrases divided with help of NPMI			
'infection'	'upper'	'tract'	'respiratory'	'infection'	'upper'	'tract'	'respiratory'
infekcja	*górnych*	*dróg*	*oddechowych*	*infekcja*	*górnych*	*dróg*	*oddechowych*
infekcja	*górnych*	*dróg*		—			
infekcja				*infekcja*			
	górne	*drogi*	*oddechowe*		*górne*	*drogi*	*oddechowe*
	górne	*drogi*		—			
		drogi	*oddechowe*			*drogi*	*oddechowe*
		drogi				*drogi*	

Table 1: The nested phrases for two methods

bigram	translation	NPMI
infekcja górna	'infection upper'	0.65658
górna droga	'upper tract'	0.78773
droga oddechowy	'tract respiratory'	0.95089

Table 2: The NPMI value for the bigrams of the phrase: *infekcja górnych dróg oddechowych*

3.3 Examples

Let us consider examples illustrating the method. We compare nested phrases obtained from the phrase *infekcja górnych dróg oddechowych* 'infection (of the) upper respiratory tract' for the two following methods: creating all grammatically correct nested phrases, and the NPMI driven method. The considered phrase is constructed according to the following pattern: Noun_j Adj_i Noun_i Adj_i where indexes indicate agreement constraints, so a grammatically correct phrase may consist of: Noun_j Adj_i Noun_i, but can't be constructed as: Noun_j Adj_i. Thus, *infekcja górnych dróg* 'infection of the upper tract' is grammatically correct, while *infekcja górnych* 'infection of upper' is not. The phrase can be divided in one of two places indicated by the 'I' character: *infekcja* I *górnych dróg* I *oddechowych*, 'infection I upper tract I respiratory'[3] and it is possible to create six grammatically correct phrases, see Table 1. Applying our method, we first count NPMI for the places of possible divisions. The NPMI value for two bigrams *infekcja górny* 'infection upper' and *droga oddechowy* 'tract respiratory' counted for the corpus described in Section 2 are given in Table 2. The lower value is for the first bigram so the phrase can be divided into: *infekcja* 'infection' and *górne drogi oddechowe* 'upper respiratory tract'. Both parts constitute nominal phrases so the phrase is divided in this place and both parts are added to the list of term candidates. In the next step only the second phrase can be recursively divided. The weaker connection is for: *górny droga* 'upper tract'. So the adjective *górna* 'upper' is cut off the phrase and only the nested phrase *drogi oddechowe* 'respiratory tract' is accepted as a term candidate. Table 1 contains all the nested phrases obtained by both methods for the considered phrase. It may be noted that our method, correctly, does not extract two semantically odd nested phrases from the six obtained by the first method.

Let us consider a phrase where the lowest NPMI indicates division into an adjective and a nominal phrase: $\textit{boczne}_{adj}$ $\textit{skrzywienie}_{noun}$ $\textit{kręgosłupa}_{noun}$ 'lateral curvature (of the) spine'. The phrase can be divided in both places: *boczne* I *skrzywienie* I *kręgosłupa* 'lateral I curvature I spine'. The weakest connection is for the bigram: *boczny skrzywienie* 'lateral curvature', it indicates division into the nominal phrase *skrzywienie kręgosłupa* 'curvature (of the) spine', and the adjective *boczne* 'lateral'. The other place of division causes the phrase to be divided into two nominal phrases. So we compare the NPMI for *skrzywienie kręgosłup* 'curvature spine', with 120% NPMI *boczny skrzywienie* 'lateral curvature', see Table 3. As the first value is lower than the second one, the method prefers to divide the phrase into two nominal phrases *boczne skrzywienie* 'lateral curvature' and *kręgosłup* 'spine'. The basic algorithm, without multiplying NPMI values in some cases by 120%, creates a good term *skrzywienie kręgosłupa* 'curvature (of the) spine' instead of two nominal phrases: *boczne skrzywienie* 'lateral curvature' and

[3]The word for word translation.

bigram	translation	NPMI	120% NPMI
boczny skrzywienie	'lateral curvature'	0.67619	0.81143
skrzywienie kręgosłup	'curvature spine'	0.80151	

Table 3: The NPMI value for the bigrams of the phrase: *boczne skrzywienie kręgosłupa*

kręgosłup spine.

There are a few cases when the phrase division driven by the NPMI value prefers cutting off an adjective in the first step instead of dividing it into two nominal phrases, see: *okołoporodowe*$_{adj}$ *uszkodzenie*$_{noun}$ *splotu*$_{noun}$ *ramiennego*$_{adj}$ *prawego*$_{adj}$ 'perinatal damage (of) right brachial plexus'. Despite the fact that *okołoporodowe uszkodzenie splotu ramiennego* 'perinatal damage (of) brachial plexus' is a good term, we would prefer the division into two nominal phrases *okołoporodowe uszkodzenie* 'perinatal damage' and *splot ramienny prawy* 'right brachial plexus'. The last division reflects the internal construction of the phrase that might be important in an ontology construction task which is one of the intended uses of the method. In this case, we want to recognise nested phrases representing two concepts which are in a relationship. The method still (correctly) cuts off the adjective *częsty* 'frequent' from the phrase *częste infekcje górnych dróg oddechowych* 'frequent infections (of the) upper respiratory tract'.

4 Terms ordering

To test to what extent our approach to the phrase selection problem influences the ultimate results of the term selection algorithm, we used the C-value coefficient (Frantzi et al., 2000) to order extracted phrases. The standard equation for this coefficient is given in (2) where p is the phrase under consideration, freq(p) is a number of occurrences of this phrase both nested and in isolation, LP is a set of phrases containing p, r(LP) – the number of different phrases in LP, and l(p) = $log_2(length(p))$.

$$C\text{-}value(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)), & if\ r(LP) > 0, \\ l(p) * freq(p), & if\ r(LP) = 0 \end{cases} \quad (2)$$

The C-value ranking method is focused on deciding which nested phrases should be considered as terms. It assigns higher values to phrases which, having the same frequency rate, occur more frequently in isolation or occur in a larger number of different longer phrases, i.e., have different lexical contexts within a set of initially extracted phrases. To account for the fact that long phrases tend to occur more rarely than shorter ones, the result is multiplied by the logarithm of the phrase length. If a phrase occurs only in isolation, its frequency rate defines the C-value. When a phrase occurs only in one context, its C-value gets the value 0 as it is properly assumed to be incomplete. If a nested phrase occurs in a lot of different contexts, its chances of constituting a domain term increase. A slight modification of the method also allows for processing phrases of length 1, which originally all got a 0 value. For this purpose, for one word phrases, the logarithm of the length (used in the original solution) is replaced with a non zero constant. In (Barrón-Cedeno et al., 2009), where this method was applied to Spanish texts, the authors set it to 1, arguing that if it is lower, one word terms are located too low on the ranking list (it cannot be greater than 1 for obvious reasons). Our experiments proved that in our data, such a change results in very many one word elements at the top of the list, we used a 0.1 value as the equivalent of logarithm of length for one word phrases.

The results obtained using the C-value method depend on the details concerning the way in which we distinguish different phrases, i.e., how we count r(LP). First, for inflectional languages like Polish, a method for recognising inflected forms of a multiword phrase has to be established. In our experiment, we used base form sequences for this purpose. Secondly, the way of counting contexts has to be elaborated. For example, it should be decided, whether *red blood cells* and *white blood cells* are two different contexts for *cell* or only one. For languages with more relaxed word order, like Polish, the same phrase can appear in different orders, e.g., *liczne krwinki białe* 'numerous white blood cells' or *krwinki białe liczne* 'white blood cells numerous'. As the C-value coefficient is drastically different for frequent phrases which occur in one and in two different contexts, we tried to limit the number of phrase types

length	all	=1	=2	3–5	>5
s-phrases	32809	4918	13442	13984	465
npmi-phrases	28328	4918	11693	11313	393
s&npmi-phrases	26671	4918	10420	10929	404

frequency	=1	2-10	11-50	51-100	101-1000	>1000
in isolation	13304	6776	1506	300	415	81
s-phrases	18572	10417	2461	523	704	132
s&npmi-phrases	15210	8296	2002	420	625	118

C-value	0	0<c<1	1≤c<5	5≤c<10	10 ≤c<100	>100
s-phrases	8946	2500	16891	1804	2312	357
s&npmi-phrases	3428	2508	16652	1672	2074	337

Table 4: The number of recognised phrases

	total	removed		lowered			
changes		all	correctly	all	incorrectly	correctly	questionable
nmpi/s-phrases	39	39	30	0	-	-	-
s&nmpi$_1$/s-phrases	137	28	26	109	19	73	17
s&nmpi/s-phrases	132	27	27	105	20	70	15

Table 5: The number of correct changes for the first 2000 positions

which differ only in order or are included one in another. We discussed different methods of counting contexts in (Marciniak and Mykowiecka, 2014) and concluded there that none of the tested ranking procedures were able to filter out all semantically odd noun phrases from the top of the list of terms. The best results we obtained taking only the nearest context of a phrase into account, i.e. the closest word to the left or to the right of a phrase. We used the greater number of these different left and right contexts. This solution can lower the actual number of contexts, but it prevents us from counting the same context words placed before and after the phrase twice.

5 Results and evaluation

We applied the C-value method to two sets of term candidates. The first set contains all possible phrases fulfilling the grammatical rules, while the second one is obtained by the method described in the previous sections. It is worth noting that we consider contexts of nested phrases only when they are recognised in phrases by the method. As both methods recognised different numbers of phrases,[4] Table 4 gives the comparison of their numbers. In this table, *s-phrases* refers to the baseline solution in which all grammatically correct nested phrases are taken into account, *npmi-phrases* refers to the solution obtained while recognising nested phrases using only NPMI value and *s&nmpi-pharses* is a name used for the final solution in which both grammar rules and NPMI values are utilised. Initially, 32809 phrases were extracted. The number of candidate phrases was significantly lower after applying NPMI selection (by 15%), but some of them were not grammatically correct. When applying both selection criteria we obtained about 80% of the phrases (only grammatically correct) from the *s-phrases* set. The reduction concerned phrases irrespective of their occurrences within texts. As to the distribution of the C-value, it may be seen that we finally obtained much fewer phrases with a 0 C-value.

In the paper (Marciniak and Mykowiecka, 2014), an evaluation of different aspects of the original C-value method applied to the same domain corpus is given. In this paper, we want to verify the tendencies

[4]The set of phrases recognised by the proposed method is included in that consisting of phrases recognised by the standard method based on all valid phrases.

of changes introduced by the proposed method. To focus on this task, we analysed all phrases that were included in the top 2000 positions ranked by the first method and whose position was moved below the 3000 in the final list, see Table 5. This comparison shows that our solution removed 6.6% (132) of phrases from the top of the list of terms, and 73.5% (97) among them were semantically odd phrases. We compared the baseline with the version in which, the minimum of NPMI value was always used to indicate phrase division (s&nmpi$_1$) and with the final version, in which the division into two noun phrases was preferred (i.e. if the NPMI at the division position was not significantly higher than the minimum inside phrase). In the first case, we observed the elimination of only 39 phrases from the top 2000. From these sequences, 9 were incorrectly removed from the candidates list. Using both NPMI value and grammaticality test resulted in 137 changes inside the top 2000. This time, from 28 removed elements only 2 could be considered correct. In the final solution, all 27 phrases eliminated form the first 2000 were correctly eliminated, while from the remaining 105 phrases, whose positions were significantly lowered, 70 were not terms. For some phrases it is difficult to judge whether they are domain related phrases or are rather related to other topics. These cases were labelled as "questionable" in the table.

As the proposed method does not change the way of counting whole phrases recognised in the corpus, we cannot expect that every incorrect phrase will be eliminated. For example, the phrase *infekcja górnych dróg* 'infection (of the) upper tract' cannot disappear from our list of term candidates, as it occurred three times as a whole phrase due to a spelling error in the word *oddechowy* 'respiratory'. We only expect that its position is similar to the position of this phrase ranked according to the frequency of the whole phrase. We obtained this required effect. The semantically odd phrase, considered above, changed its position from 144 to 4374.

The presented results show that integrating NPMI with syntactic rules resulted both in better selection and ranking of candidates. The final decision to prefer division into two noun phrases had rather small but positive effects.

6 Conclusion

In the paper, we described a method for recognising nested phrases based on normalised pointwise mutual information. We proved that the method has a strong tendency not to recognise semantically odd phrases once they are nested, and allows for the elimination of incorrect unfinished phrases from the top part of the ranking list. The method can be applied to any language: it requires the existence of a POS tagger and several rules describing noun phrase structure. Taking into account information on the internal syntactic structure of terms improved the results.

There are several possible directions for further research. First, we plan to test the method on different datasets. Then, some extensions of the method are planned. The potentially easiest one concerns the problem of how to extend the method to take into account more complex phrases (i.e. prepositional phrases and coordinated phrases) and count NPMI effectively for them. The second problem refers to longer phrases that are strongly connected but only when all elements appear together. An example of such a phrase is *wykładnik stanu zapalnego* 'inflammation exponent' where *stan zapalny* 'inflammation' can appear in different contexts, but *wykładnik stanu* 'exponent (of the) state' implies the word *zapalny* 'inflammatory'. The third problem is to explore whether the proposed method provides a good starting point for recognising pieces of information that should be represented in a domain ontology.

References

Szymon Acedański. 2010. A morphosyntactic Brill tagger for inflectional languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Alberto Barrón-Cedeno, Gerardo Sierra, Patrick Drouin, and Sophia Ananiadou. 2009. An improved automatic term recognition method for Spanish. In *Computational Linguistics and Intelligent Text Processing, LNCS 5449*, pages 125–136. Springer.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to*

Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, volume Normalized, pages 31–40, Tübingen.

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.

Kyo Kageura and Bin Umino. 1996. Method for automatic term recognition. A review. *Terminology*, 3:2:259–289.

Ioannis Korkontzelos, Ioannis P. Klapaftis, and Suresh Manandhar. 2008. Reviewing and evaluating automatic term recognition techniques. In *Advances in Natural Language Processing, LNAI 5221*, volume 5221, pages 248–259. Springer.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2011. Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. In *Proceedings of BioNLP 2011*, pages 92–100.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2013. Terminology extraction from domain texts in Polish. In R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform. Advanced Architectures and Solutions*, volume 467 of *Studies in Computational Intelligence*, pages 171–185. Springer-Verlag.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2014. Terminology extraction from medical texts in polish. *Journal of Biomedical Semantics*, 5:24.

Patrick Pantel and Dekang Lin. 2001. A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 36–46, London, UK, UK. Springer-Verlag.

Maria T. Pazienza, Marco Pennacchiotti, and Fabio M. Zanzotto. 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In S. Sirmakessis, editor, *Knowledge Mining Series: Studies in Fuzziness and Soft Computing*. Springer Verlag.

Francesco Sclano and Paola Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In Ricardo Jardim-Gonçalves, Jörg P. Müller, Kai Mertins, and Martin Zelm, editors, *Enterprise Interoperability II*, pages 287–290. Springer.

Juan A. Lossio Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2014. Towards a mixed approach to extract biomedical terms from documents. *International Journal of Knowledge Discovery in Bioinformatics*, 4(1).

Thuy Vu, Ai Ti Aw, and Min Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of International Joint Conference on Natural Language Processing*.

Joachim Wermter and Udo Hahn. 2005. Massive biomedical term discovery. In *Discovery Science, LNCS 3735*, pages 281–293. Springer Verlag.

Marcin Woliński. 2006. Morfeusz — a practical solution for the morphological analysis of Polish. In *Intelligent Information Processing and Web Mining. Proceedings of the International IIS:IIPWM'06 Conference held in Ustron, Poland*. Springer-Verlag.