Nested term recognition driven by word connection strength

Małgorzata Marciniak and Agnieszka Mykowiecka

Institute of Computer Science, Polish Academy of Sciences

Jana Kazimierza 5

01-248 Warsaw, Poland

mm@ipipan.waw.pl, agn@ipipan.waw.pl

Abstract

Domain corpora are often not very voluminous and even important terms can occur in them, not as isolated maximal phrases but only within more complex constructions. Appropriate recognition of nested terms can thus influence the content of the extracted candidate term list and its order. We propose a new method for identifying nested terms based on a combination of two aspects: grammatical correctness and normalised pointwise mutual information (NPMI) counted for all bigrams in a given corpus. NPMI is typically used for recognition of strong word connections, but in our solution we use it to recognise the weakest points to suggest the best place for division of a phrase into two parts. By creating, at most, two nested phrases in each step, we introduce a binary term structure. We test the impact of the proposed method applied, together with the C-value ranking method, to the automatic term recognition task performed on three corpora, two in Polish and one in English.

Keywords

automatic term extraction, domain corpora, nested phrase recognition, pointwise mutual information, C-value

1. Introduction

The Automatic Term Recognition (ATR) task consists in identifying linguistic expressions that refer to domain concepts. This process is usually carried out 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 & 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. 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 & Schutze, 1999), while ATR applications based on them are described in e.g., (Pantel & Lin, 2001), (Sclano & Velardi, 2007). A comparison of these approaches is given in (Pazienza, Pennacchiotti, & Zanzotto, 2005). Some hybrid solutions to the ATR problem have also been proposed in (Vu, Aw, & Zhang, 2008) or (Ventura, Jonquet, Roche, & Teisseire, 2014). In the paper (Korkontzelos, Klapaftis, & Manandhar, 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. So, in our research on ATR in Polish, we accepted the most widely recognised termhood-based method as a starting point, that is the C-value method, (Frantzi, Ananiadou, & Mima, 2000).

An important feature of this method that attracted our attention was the focus on nested terms. Frantzi, Ananiadou and Mima (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 term extraction, but they also give examples where a nested phrase constructed according to the grammar rules is not a term. One of 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.

Although a lot of different candidate term ordering procedures have been proposed in related studies, none of them allow for clear separation of domain-related terms from non-related ones or properly structured terms from accidental word groups. In our paper, we decided to focus on the first phase of the term extraction procedure, that is on candidate selection, especially in nested term recognition. Improvements made at this stage can eliminate some improperly structured candidate terms, making the final results better.

We applied the C-value method to extract terminology from a corpus of hospital discharge documents in Polish. Experiments, in which different methods of counting the C-value were tested, are described in (Marciniak & 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 appearing among the 200 top positions are:

- *USG jamy* ‘USG of cavity’ being a nested fragment of the very frequent phrase *USG jamy brzusznej* (USG cavity abdominal)[1] ‘USG of abdominal cavity’,
- *infekcja górnych dróg* ‘infection of upper tract’ is a part of *infekcja górnych dróg oddechowych* (infection upper tract respiratory) ‘infection of the upper respiratory tract’,
- *powiększony węzeł* ‘enlarged node’ is a part of *powiększony węzeł chłonny* (enlarged node lymph) ‘enlarged lymph node’.

We observed that semantically odd phrases are created by the truncation of a word (or a phrase) from a semantically correct phrase in which the truncated part is rather strongly connected to its preceding word, e.g. *jama* ‘cavity’ and *brzuszna* ‘abdominal’ create a strong collocation that is part of many other terms, such as: *tomografia jamy brzusznej* ‘tomography abdominal cavity’, *narządy jamy brzusznej* ‘organs of abdominal cavity’ or *badanie jamy brzusznej* ‘examination of abdominal cavity’. All these words: tomography, examination, and organs might be connected with many other phrases, but *jama* ‘cavity’ is almost only connected with *brzuszna* ‘abdominal’ or *ustna* ‘oral’ in the data.

We proposed a method that prevents the creation and promotion of such truncated phrases to be considered as terms. The main idea was to use a unithood-based method e.g., Normalised Pointwise Mutual Information (NPMI) (Bouma, 2009) for driving recognition of nested phrases. Our solution was 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 or one such phrase and a modifier. Usually, there are several possible places for division of a phrase. From all of them, we chose the weakest point according to the NPMI counted for bigrams on the basis of the whole corpus. So, as a bigram constitutes a strong

[1] In the paper, the word for word translation is given in parenthesis.

collocation, it prevents the phrase from being divided in this place, and does not usually lead to the creation of semantically odd nested phrases, of which examples are given above.

We tested the ideas presented above on two datasets in Polish (medical texts and economic corpus) and one in English, i.e. GENIA (Kim, Ohta, Tateisi, & Tsujii, 2003). They are described in Section 2. In the next section, we suggest a way of recognising terminology phrases, and discuss differences in the structure of such phrases in Polish and English. In the following two sections, we present the method in detail. Then, in Section 6, we describe a 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. Moreover, for economic data we present an evaluation of the top 1K terms obtained by the traditional C-value method and the method proposed in the paper. Finally, we discuss the impact of the method on recognising terms in the GENIA corpus.

2. Description of domain corpora

The experiments were performed on three domain corpora: two in Polish and one in English. The first Polish corpus contains hospital documents which are not publicly available, while the second one – plWikiEcono – contains economic articles and is available from http://zil.ipipan.waw.pl/plWikiEcono. The experiment on the English data was performed on the well-known GENIA corpus publicly available from http://www.nactem.ac.uk/genia/. It contains Medline abstracts of articles concerning molecular biology and is available together with annotations made at various levels. The data is annotated with part of speech and biological terms, among others.

The Polish medical 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. Most information is given as free-form text but the data also contains a lot of test results with numerical values. The structure of discharge records is fixed; they consist of

the same parts describing: patients, test results, diagnoses and recommendations. As patients from a ward often suffer from similar diseases, the scope of words used in the documents is limited and many phrases are repeated within the documents.

The economic corpus contains 1219 articles from Wikipedia. It consists of the textual content of articles that have economy related headings and those linked to them. This data contains encyclopaedic information, so each document concerns one topic and introduces a new vocabulary related to it.

The collected texts were analysed using standard general purpose NLP tools. The morphological tagger, Pantera (Acedański, 2010), cooperating with the Morfeusz analyser (Woliński, 2006), was used to divide the text into tokens and annotate them with morphosyntactic tags. They included a base form and a part of speech name (POS), as well as case, gender and number information, where appropriate. This information is used by shallow grammars recognising the boundaries of nominal phrases − term candidates and, also, sources for nested phrases. We call them maximal phrases hereinafter.

As the automatic tagging of medical documents is difficult (Marciniak & Mykowiecka, 2011), the annotation was partially corrected by a set of global rules. They did not take context into consideration and they were used only to eliminate some systematic errors (replace very unlikely interpretations of homonyms) and to introduce interpretation of the most common abbreviations or units. Moreover, some improperly recognised sentence endings after abbreviations were removed. In the case of economic texts, we defined an additional domain dictionary containing 741 entries of unknown word-forms. To correct some Pantera decisions, we defined a set of 156 context rules in Spejd (Przepiórkowski, 2008), which corrected 1.5% of word-form descriptions. For example, one of the rules corrects the interpretation of *U* in all occurrences of the string *Dz.U.* that abbreviates the phrase *Dziennik Ustaw* ‘Journal of Law’. *U* is interpreted by the analyser as the preposition ‘at’, but in this context it should be interpreted as the abbreviation of the word *ustawa* ‘law’.

	Medical data	plWikiEcono
all tokens	2036068	474681
numbers	252568	14827
punctuation marks	846778	94725
words	888883	353732
unrecognised strings	47839	11397

Table 1: Corpora statistics

Table 1 gives the statistics of token types recognised in both corpora. The texts contain quite a lot of words unrecognised by Morfeusz. In the case of the medical corpus, the reason is twofold: the vocabulary of clinical documents significantly differs from general Polish texts and the texts are not very well edited despite the spelling correction tools usually being turned on, so they contain a significant number of misspelled words. This results in 48,000 unrecognised tokens, many of them are medications, diagnoses written in Latin, and abbreviations. In the economic corpus, unrecognised words are mainly proper names that are not included in the general dictionary and English equivalents of concepts described in Polish Wikipedia articles. Only tokens classified as words can be elements of terms. A large number of unrecognised strings reduces the number of phrases and affects the quality of some of them. It is worth noting that, although the medical corpus has 4.2 times more tokens than the economic one, the number of words in the medical data is only 2.5 greater than in the latter.

The GENIA corpus is a collection of Medline 1999 abstracts containing approximately 500,000 tokens. The texts are very different from everyday English as they contain a lot of proper names, abbreviations, chemical and numerical expressions, a lot of strings with hyphens and slashes, which make tokenisation difficult. Therefore, we decided to use the POS annotation available together with the corpus, as it was corrected by human annotators (Tateisi & Tsujii, 2004). This annotation consists of tokens and POS assigned to them, but does not contain lemmas. As our method proposed in the paper uses lemmas (it works better on lemmas), we decided to add them to the GENIA annotation. We analysed GENIA texts with the Stanford

tagger (Toutanova, Klein, Manning, & Singer, 2003), and used the results as a source of lemmas. For each pair of word form and its POS from GENIA's original annotation, we looked for the automatic annotation done with the Stanford tagger. If the pair existed, we used the lemma assigned by the Stanford tagger. If not, we searched for the word form with any POS assigned and, if that failed too, we treated the word form as the lemma.

3. Noun phrase extraction

Maximal noun phrases extracted from texts with respect to linguistic rules are the most commonly used source of term candidates. For Polish, grammatical constraints resulting from the case and gender and number agreement allow us to significantly reduce the number of considered phrases to syntactically valid ones. In this section, we describe the construction of Polish noun phrases, compare them to their English equivalents, and describe shallow grammars used to identify maximal noun phrases within the texts.

In the extraction step, we identify complex noun phrases consisting of nouns with adjectival and nominal modifiers obeying language dependent grammar rules. For Polish, the rules represent, in particular, case, gender and number agreement. For the task of terminology extraction, the types of Noun Phrases (NP) under consideration can be limited to those schematically defined as below:

$$\text{NounPhrase}_k \rightarrow \text{Noun}_k$$

$$\text{NounPhrase}_k \rightarrow (\text{AdjPhrase}_k)^* \ \text{NounPhrase}_k \ (\text{AdjPhrase}_k)^*$$

$$\text{NounPhrase}_k \rightarrow (\text{AdjPhrase}_k)^* \ \text{NounPhrase}_k \ \text{NounPhrase}_{gen}$$

In Polish noun phrases, adjectives can occur at both sides of the noun. Adjectives preceding nouns usually define some features of a following noun phrase, while adjectives placed after a noun have a classification role, e.g.: *pilna kontrola neurologiczna* (urgent control neurological) 'urgent neurological control', where the adjective 'neurological', following the noun, classifies

the type of ‘control’, while ‘urgent’ describes a feature of the ‘control’. This is only a general rule as, for example, in *echogeniczność miąższu prawidłowa* (echogenicity parenchymal normal) ‘normal parenchymal echogenicity’ the adjective ‘normal’ describing the feature of NP occurs after it. The phrase appears in the medical texts 34 times, while the more typical phrase: *prawidłowa echogeniczność miąższu* occurs only 23 times. The free order of Polish adjectives caused problems that are not present in English, where all adjectives precede the modified noun. So, in English, if an adjective occurs between two nouns it always modifies the second noun.

Polish phrases constructed according to the third rule above (the sequence of two Noun Phrases: NP1 $NP2_{gen}$) refer in English to a phrase NP2 NP1 or (less often) to a phrase with the preposition ‘of’. For example, *wynik badania* (result examination) refers to English ‘examination result’, but also ‘result of examination’. In Polish, a noun phrase sequence might be quite long, e.g.: *wodonercze niewielkiego stopnia dolnego układu podwójnego nerki prawej* ‘mild hydronephrosis of the duplicated lower collecting system of the right kidney’. In English, such long sequences of NP phrases contain the preposition ‘of’. In GENIA, almost 75,000 occurrences of almost 35,000 phrases are annotated for biological terms. But only 255 occurrences of phrases (of 193 types) contain the preposition ‘of’. In all of the data, there are more than 21,000 occurrences of the preposition ‘of’ and, in the vast majority of cases, such phrases are annotated as two separate terms e.g.: <monocyte-specific function> of the <peri-kappa B factor>. Because adding prepositional phrases to the grammar will result in many thousands of additionally recognised phrases, only a small proportion of which are annotated with GENIA terms, we decided not to take ‘of’ phrase modifications into account.

The examples of GENIA nominal terms cited below show that, in this case, a lot of term elements are numbers and some punctuation marks also have to be considered as term elements:

- "octamer" motif,
- human immunodeficiency virus type 1 (HIV-1) infection,

- Ca2+ /calmodulin-dependent protein phosphatase,
- Cushing's patients,
- TH1 clone 29,
- -294 to -251 bp,
- in vitro polyclonal B cell immunoglobulin (Ig) response,
- "B-subunit knock-out" (BKO) construct,
- AP-1, NF-AT, and NF-kB motifs,
- IL-2 and the IL-2-R alpha gene.

Most of the cited example elements are tagged in GENIA with JJ or NN tags so their recognition using POS data is easy, but it is not always the case. For example, *29* in *TH1 clone 29* or *II* in *class II* are tagged as numbers and "*B-subunit knock-out" (BKO) construct* is tokenised as a sequence of "JJ JJ " (NN) NN tags. The grammar rules are further complicated by the fact that determiners can occur inside GENIA complex phrases but do not begin them. As some coordinated terms are annotated jointly (the last two examples), we added a rule for coordination which results in the wrong recognition of phrases which, although they occur within coordinated phrases, were annotated separately. Several problems concerning GENIA terms containing prepositions and those encoded within coordination are described in (Nenadic, Spasic, & Ananiadou, 2005).

The general schema of the shallow English noun phrase grammar defined for the task is as follows (GENIA (Stanford) POS tags names are used):

Noun → NN | NS | FW

Adjective → JJ | NN POS | RB ADJ | VBN | VBG

Numeral → CD | CD POS | CD %

AdjPhrase → Adjective | Noun-Adjective | Noun POS

NounP → Noun | Numeral

NounP → AdjPhrase * NounP

NounPhrase → NounP CC NounP

NounPhrase → NounP DT NounP

NounPhrase → NounP

4. Nested phrase recognition

In this section, we describe how to create a list of term candidates that takes into account nested phrases. Nested phrases are syntactically valid phrases included in maximal noun phrases. The idea to use the NPMI method for limiting the recognition of nested terms to semantically valid ones was inspired by the analysis of Polish clinical texts. The first data set it was tested on was the corpus of Polish hospital documents, so we explain it using examples from these texts.

4.1. Motivation

The original C-value method (Frantzi, Ananiadou, & Mima, 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', which is created from the frequently occurring phrase (126 occurences as a maximal phrase, and 44 as a nested one) *infekcja górnych dróg oddechowych* (infection upper tract respiratory), ‘infection (of the) upper respiratory tract’. The last phrase has many different longer phrases in which it is nested, e.g.: (*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 observed 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 subcategorisation frames, which do not usually fulfil 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 (NP NP_{gen}). 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 this phrase should be divided into two nominal phrases: *prawidłowa mikroflora* 'normal microflora' and *górne drogi oddechowe* 'upper respiratory tract'. To account for this observation, we may slightly prefer divisions into two nominal phrases instead of an adjective and a nominal phrase. However, this 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. In this case, the division should be made after the first adjective. To account for this observation, in a case when division into two noun phrases is possible but is not strong enough, we prefer to divide a phrase into modifier and noun phrase parts.

4.2. **Simplified base phrases**

Polish is a highly inflected language, so to identify and compare nested phrases 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 & 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. This approach also allows for easier recognition of a nested term *gardło* inside the complex term *zapalenie gardła* as we compare it to the simplified form of *zapalenie gardło*.[2] This approach is much simpler and more effective than comparing formally correct phrases, but very rarely do semantically different phrases have the same simplified base forms, and problems concerning this approach are discussed in (Marciniak & Mykowiecka, 2014).

4.3. **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, would be high if the rare token only appeared 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 tokens, is given below. In the equation, *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.

$$NPMI(x,y) = \ln\frac{p(x,y)}{p(x)*p(y)} / -ln(p(x,y)) \quad (1)$$

First, we extracted all the grammatical phrases from the corpus, taking into account only the maximal ones. Then, for each phrase we identified all places where it can be divided according to grammar rules. We counted NPMI for those division sites and indicated the weakest connection in the phrase. Then, we divided the phrase into two parts in this position and increased the embedded occurrences counted for those fragments which are nominal phrases recognised by our grammar. The pseudocode representing this schema is given below.

[2] 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).

```
nested_phrases(phr)
 if phr is a valid noun phrase add phr  to the list of terms
 if length(phr)>1
    find all i positions  where phr can be divided according to the grammar 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
    divide phr into phr1 and phr2 on j-th position
    nested_phrases(phr1)
    nested_phrases(phr2)
```

Figure 1 Initial procedure of nested phrases recognition

Experiments performed using the above schema for Polish medical term extraction showed that, in some cases, the two lowest NPMI values are not very different, and the correct place of division is indicated not by the lowest NPMI value, but by the second one. This can be either the result of a small number of different modifiers or the data size being too small to provide enough examples. To account for this observation, we decided to introduce a linguistically motivated heuristic. The specificity of adjectives in Polish nominal phrases described in 4.1 results in the observation that modifiers which occur at the end of the phrase are more strongly related to the modified noun (in Polish, specifying adjectives occur mostly after nouns) than those which occur at the beginning (these are often attributive expressions). Thus, we accepted the division in the place indicated by the weakest NPMI value if it resulted in two grammatical nominal phrases or in one element followed by a grammatical nominal phrase. Otherwise, we found the weakest place where the phrase was divided into two nominal phrases. Then, we compared the NPMI value referring to the bigram occurring in the division place with 120% (fixed experimentally) of the lowest NPMI value. If this was the case, we chose the second best dividing position, if not, we divided the phrase at the NPMI minimum. In any case, both fragments were further divided until we got one word fragments. The modified schema is shown in Figure 2.

```
nested_phrases(phr)
 if phr is a valid noun phrase add phr  to the list of terms
 if length(phr)>1
    find all i positions  where phr can be divided according to the grammar 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 or into a one-word segment and
                                        a nominal phrase
      divide phr into phr1 and phr2 on j-th position
    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
      else
        divide phr into phr1 and phr2 on j-th position

     nested_phrases(phr1)
     nested_phrases(phr2)
```

Figure 2 Modified procedure of nested phrases recognition

4.4. **Examples**

Let us consider some examples that illustrate the method. We compared 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 ‘|’ character: *infekcja* | *górnych dróg* | *oddechowych,* (infection | upper tract | respiratory) and it is possible to create six grammatically correct phrases, see Table 2. Applying our method, we first counted NPMI for the places of possible divisions. The NPMI value for two bigrams *infekcja górnych* ‘infection upper’ and *dróg oddechowych* ‘tract respiratory’ counted for the medical

corpus of Polish texts described in Section 6 are given in Table 3. 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: *górne drogi* ‘upper tract’. So the adjective *górne* ‘upper’ is cut from the phrase, and only the nested phrase *drogi oddechowe* ‘respiratory tract’ is accepted as a term candidate. Table 2 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.

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 2:The nested phrases for two methods

bigram	translation		NPMI
infekcja górnych	‘infection	upper’	0.65658
górnych dróg	‘upper	tract’	0.78773
dróg oddechowych	‘tract	respiratory’	0.95089

Table 3: The NPMI value for the bigrams of the phrase: *infekcja górnych dróg*

Let us consider a phrase where the lowest NPMI indicates division into an adjective and a nominal phrase: $boczne_{adj}$ $skrzywienie_{noun}$ $kręgosłupa_{noun}$ ‘lateral curvature (of the) spine’. The phrase can be divided in both places: *boczne* | *skrzywienie* | *kręgosłupa* ‘lateral | curvature | spine’. The weakest connection is for the bigram: *boczne 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 kręgosłupa* ‘curvature spine’, with 120% NPMI *boczne* skrzywienie ‘lateral curvature’, see Table 4. 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 *kręgosłup* spine.

bigram	translation	NPMI	120% NPMI
boczne skrzywienie	‘lateral curvature’	0.67619	0.81143
skrzywienie kręgosłupa	'curvature spine’	0.80151	

Table 4: The NPMI value for the bigrams of the phrase: *boczne skrzywienie kręgosłupa*

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’.

An example of how our algorithm divides English phrases is given in Figure 3 below. The phrase is not annotated in GENIA by itself but it consists of a coordination of two annotated phrases: <*cons lex="activated_human_peripheral_blood_lymphocyte" sem="G#cell_type"* >activated human peripheral blood lymphocytes</*cons*> and <*cons lex="AA2_cell" sem= "G#cell_line"*> AA2 cells</*cons*>.

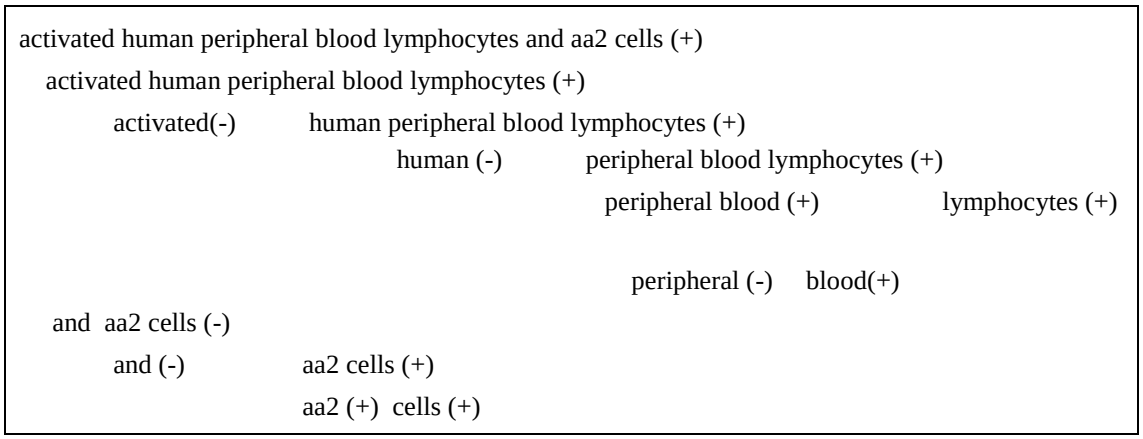

Figure 3: The NPMI driven division of an English biomedical phrase

The phrases marked with the plus sign in Figure 3 are recognised as nested terms by our algorithm. From these, five phrases i.e. <*human peripheral blood lymphocytes*>, <*peripheral blood lymphocytes*>, <*peripheral blood*>, <*lymphocytes*> and <*AA2 cell*> are also identified as terms within the GENIA corpus (in other text fragments). We additionally postulated <*AA2*> and <*cells*> terms, which are not annotated by themselves anywhere in GENIA texts. An existing GENIA term <*human peripheral blood*> is not recognised by the method inside this phrase, but was recognised elsewhere.

5. Terms ordering

In order 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, Ananiadou, & Mima, 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 the number of occurrences of this phrase both nested and in isolation, and *LP* is a set of phrases containing *p*, *r(LP)* – the number of different phrases in *LP, and* $l(p) = log_2(length(p))$.

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

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, Sierra, Drouin, & Ananiadou, 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 the 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 cell*s are two different contexts for *cell* or only one. For languages with a 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 which differ only in order or are included one in another. Different methods for counting contexts are described in (Marciniak & Mykowiecka, 2014); the authors 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 were obtained by 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 reduce the actual number of contexts, but it prevents us from counting the same context words placed before and after the phrase twice.

6. Results and evaluation

6.1. **Polish medical data**

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 the procedure may result in the same

sequence of words being recognised as a nested phrase in some contexts and not being recognised as such in different ones. Only contexts for the recognised nested terms are used to calculate the C-value.. The two tested methods recognised different numbers of phrases.[3] In Table 5, which gives a comparison of these numbers, *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 values, and *s&nmpi-pharses* is the 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 ones) from the *s-phrases* set. The reduction targeted phrases irrespective of the number of their occurrences within texts.

For the *s-phrases* candidate set we applied the standard C-value candidate ordering schema, using two strategies for identifying numbers of different contexts. The first one, named as *s-phrases-all*, consisted in counting all different contexts a particular phrase occurred in, while in the s-*phrases-max-sg* variant, we took the maximal value from the number of left and right contexts counted separately. These strategies resulted in different distributions of the C-value. It can be seen that, in the second approach, we obtained more phrases with lower C-values. The analysis of the results showed that most of the phrases which obtained a 0 C-value using the *s-phrases-maxsg* method were correctly eliminated,; so, for the *npmi* variant we chose this strategy. For the *s&npmi* set, the overall number of phrases was much smaller but the distribution of non-zero C-values remained almost the same, as we obtained much fewer phrases with a 0 C-value.

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

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

C-value	0	0<c<1	1<=c<5	5<=c<10	10<=c<100	>100
s-phrases-all	5318	3330	19214	2060	2514	373
s-phrases-maxsg	8946	2500	16891	1804	2312	357
s&npmi-phrases	3428	2508	16652	1672	2074	337

Table 5 The number of recognized 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 6 The number of correct changes for the first 2000 positions

In the paper, (Marciniak & Mykowiecka, 2014), an evaluation of different aspects of the original C-value method applied to the same domain corpus is provided. In this work, we wanted to verify the tendencies 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 whose position was moved below the 3000 in the final list, see Table 6**Błąd! Nie można odnaleźć źródła odwołania.**. 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 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 a 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 version, 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 Table 6.

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.

6.2. **Economic data**

For the economic data, we evaluated 1K terms from the top of the ranking list created according to two methods: the original C-value method, and with the help of the NPMI-based nested term recognition. These texts have an encyclopaedic character. Each document introduces new terms, but their frequencies in the data are not high. A term, to which an article is devoted, is quite often mentioned only a few times and does not appear in other articles. Texts also contain information which does not directly concern the economy, e.g. historical background. The list of term candidates (constructed without NPMI modification) consists of more than 70,000 terms,

so it is more than two times longer than for medical data, while the size of medical texts is more than two times greater in terms of recognised words. The C-value of the top part of the ranking list is lower in the economic data, as the first term *papier wartościowy* 'security' has C-value 279.92 and a similar C-value in the medical data belongs to a term placed in the 130th position. The list of term candidates for the C-value method, together with the NPMI nested term recognition, consists of almost 55,000 terms, so the method eliminates about 15,000 terms.

Table 5 shows how many economic, general and invalid terms are recognised in the tested terms. A term is classified as an invalid one if it does not have any sense in any domain (it is incomplete, wrongly structured etc.). As a general term, we classify phrases that might be considered as terms in other domains or are common in everyday language, like *punkt widzenia* 'point of view', *przykład taki* 'such example' and many one-word phrases that are difficult to evaluate. One-word phrases are always correct terms, according to our definition, but quite often have different meanings in various domains. There are one-word phrases that are clearly economic ones, like: *kredyt* 'credit' *cło* 'duty' or *podatek* 'tax'. But the decision regarding *droga* 'a way', which may be part of the phrase *budowa dróg* 'road construction' or may mean 'a method' e.g. in the phrase *drogą nabycia akcji* 'through the acquisition of shares', is not obvious. In such cases, we accepted the phrase as an economic term. Phrases like: *character* 'character' or *chwila* 'a while', *człowiek* 'a man' are classified as general terms.

	economic	general	invalid
C-value *(s-phrases*)	860	127	13
C-value and NPMI (*s&npmi-phrases*)	883	112	5

Table 7. Evaluation of the 1000 economic terms

The data included in Table 5 shows that, in this case, our method of counting nested phrases improves results a little (less than for medical corpus). The number of invalid phrases drops from 1.3 % to 0.5%. Our method removes seven truncated phrases like *giełda papierów* created from *giełda papierów wartościowych* 'stock exchange' or *spółka prawa* created from *spółka*

prawa handlowego ‘commercial law company’ or *spółka prawa cywilnego* ‘civil law partnership’. The five invalid phrases remaining in the data contain nouns that require complements. Most of them should contain prepositions that were not considered in the grammar used for initial phrase selection. For example, the phrase *ryzyko związane* is part of the phrase *ryzyko związane z (czymś)* ‘risk associated with (something)’. The percentage of economic phrases recognised by our method increased slightly from 86% to 88.3%.

6.3. GENIA **corpus**

To test whether our method can be efficient for term extraction from texts in other (somehow related) languages, we applied it to identify English biomedical terms annotated in the GENIA corpus. The original GENIA annotation covers more than 33850 different terms which occurred about 80000 times (only about 7800 terms occurred more than once). Our grammar (described in Section 3) recognised 47481 maximal nominal phrases as term candidates. The syntactical approach resulted in identifying 26590 nested phrases while NMPI based approach reduced this number to 15660. 56% of these phrases were the same as those annotated in GENIA while 2160 (6%) of GENIA phrases were not identified by our grammar. Some of phrases that weren’t annotated in GENIA might be fragments of more complex terms, some are not terms at all, while many of them were just too general for being annotated in such a specialised corpus (e.g. *cell*). Within this work we did not apply any methods for separating domain related terminology from general language expressions but we focused on eliminating term candidates which are badly structured (syntactically or semantically). Such candidates are a direct consequence of taking into account nested terms which never occur outside other term contexts. To compare syntactic (*s-phrases-maxsg*) and NPMI based strategies for nested term identification, we checked the beginning of the lists obtained by these two methods. The first 200 multiword terms from these lists were searched for on the list of GENIA terms and checked for their correctness. The results are presented in Table 8. The main reason for the difference between original GENIA terms and terms which we judged to be correct are terms in which the

full name of a term is followed by its acronym, e.g. *activator protein-1 (AP-1)*. We considered such phrases as one term, while in GENIA two separate annotations were introduced, e.g. <cons lex="interleukin-10" sem="G#protein_molecule">*Interleukin-10*</cons> *(*<cons lex="IL-10" sem="G#protein_molecule"> *IL-10*</cons>*), like* ….

	s-npmi		*s-phrase-maxsg*		examples
	number	ratio	number	ratio	
GENIA terms	157	0.785	152	0.76	tumor necrosis factor
extension of GENIA terms	19	0.095	10	0.05	Interleukin-2 (IL-2)
correct not annotated terms	1	0.01	3	0.015	major histocompatibility complex
coordination of terms	0	0.005	2	0.01	mRna and protein
Correct in total	177	**0.885**	167	**0.835**	
incorrect terms	13	0.065	28	0.14	B (nf-kappa b)
general phrases	10	0.05	5	0.025	site, high level

Table 8 Evaluation of terms extracted from GENIA corpus

The exact matching for GENIA terms (with only singular and plural numbers unified) provided slightly better results for the NPMI based method — the precision for the top 200 multiword terms was equal to P_1@200=0.785. Detailed manual evaluation showed that introducing NPMI based phrase divisions resulted in obtaining fewer incorrect terms (6.5% in place of 14%), while more (19 terms in place of 10) occurred as proper GENIA terms followed by their correct abbreviations. Taking all correct terms into account, we got P_2@200=0.885 for the NPMI based method in comparison to P_2@200=0.835 for the standard approach. In Table 9 we show precision, counted in the same way as in Table 8, for the first 100, 200, 500 and 1000 terms and multiword terms for the syntactic and NPMI based approaches.

	s-phrases-maxsg						s-phrases/npmi					
	all terms			multiword terms			all terms			multiword terms		
	P_1	P_2	%general	P_1	P_2	%general	P_1	P_2	%general	P_1	P_2	%general
100	.82	.83	.01	.82	.83	.02	.76	.82	.11	.81	.88	.03
200	.75	.82	.03	.76	.82	.03	.76	.82	.11	.79	.89	.07
500	.71	.80	.08	.76	.81	.05	.72	.83	.12	.75	.88	.08
1000	.71	.79	.08	.71	.80	.06	.72	.82	.13	.73	.85	.09

Table 9 Automatic and manual evaluation of 1000 top terms

A direct comparison of results from other terminology extraction experiments performed on the GENIA corpus is difficult, as the set of candidate terms in every approach is usually different, and quite often the ranking procedure is not described precisely enough. However, Lossio-Ventura, Jonquet, Roche, and Teisseire (2014), for example, listed four improper terms which were identified among the top-10 ranked 3-gram terms using their term ranking methods.. The first one (*kappa b alpha*) was not eliminated even by the best approach they proposed. The positions of these terms on our list are shown in Table 10.

	position in s-phrases	position in s-npmi-phrases
kappa b alpha	33	25702
virus type 1	*not present*	*not present*
c-fos and c-jun[4]	190	312
transcription factor nf-kappa	*not present*	*not present*

Table 10 Exemplary differences in phrase ordering

Even from these few examples, it can be seen that one of the most difficult cases to filter out is incomplete phrases (three from the cited four examples are invalid because they lack important modifiers). As our method already successfully eliminated quite a large number of such badly constructed phrases at the candidate selection phase, our results are better, even when we used the standard C-value term ordering method.

Our baseline results, shown in Table 8, were already better than those presented in (Lossio-Ventura, Jonquet, Roche, & Teisseire, 2014) who reported P@200 equal to 0.69 for C-value, 0.7715 for F-TFIDF-C and 0.77 for LIDF-value. In Table 11 we show the comparison of some of their results with ours (only multiword terms are taken into account.). In this table, P_1 represents exact term matching as before. In P_3 we also judge as correct, terms which are followed by their acronyms in parenthesis (like in P_2 defined for Table 9) but we did not check all the results manually as in P_2. The numbers presented in Table 11 show that our list contains

[4] This expression consists of two coordinated valid terms and thus can be also treated as a valid one.

fewer improper sequences among the top 5000 terms. At position 10000, the results achieved by (Lossio-Ventura, Jonquet, Roche, & Teisseire, 2014) are better, but these results are not very reliable. There are only about 33000 different terms annotated in the GENIA corpus, so this part of the list already contains sequences of very low frequency. Moreover, terms annotated within GENIA are not very long (the average length is equal to 3.8). A simple change within the C-value definition, aimed at reducing the weight assigned to longer terms (changing log2 into log4(length)), results in P_1@10000 equal to 0.48 and P_3@10000 equal to 0 .51 for the NPMI-based method. The additional problem is the fact that manual annotation of the GENIA corpus is not very consistent, so some phrases which are assumed to be incorrect may be, in fact, correct ones.

	c-value	F-TFIDF-C	LIDF-value	s-phrases-maxsg		s-phrases-npmi	
	P	P	P	P_1	P_3	P_1	P_3
1000	.615	.618	.697	.703	.730	**.729**	.769
2000	.570	.557	.662	.670	.695	**.707**	737
5000	.498	.482	.575	**.595**	.614	.592	.613
10000	.428	.412	**.526**	.441	.464	.414	.442

Table 11 Automatic evaluation of 10000 top terms

7. **Conclusion**

In the paper, we have described a method for recognising nested phrases based on normalised pointwise mutual information. We applied it to three corpora of various types and languages. The method can be applied to any language: it requires the existence of a POS tagger and several rules describing noun phrase structure. For all the corpora, 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 efficiency of the method depends on the corpus features: their size, thematic homogeneity

and the frequency of phrases. It would be interesting to investigate features that highlight corpora for which the method gives a significant improvement in the results.

There are several possible directions for further research. Some extensions of the method are planned for counting NPMI effectively for more complex phrases i.e. prepositional phrases and coordinated phrases. The most interesting question 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.

8. **Bibliography**

Acedański, Szymon. "A morphosyntactic Brill tagger for inflectional languages." Edited by Hrafn Loftsson, Eirikur Rognvaldsson and Sigrun Helgadottir. *Advances in Natural Language Processing.* Springer, 2010. 3-14.

Adam Przepiórkowski. *Powierzchniowe przetwarzanie języka polskiego.* Akademicka Oficyna Wydawnicza EXIT, 2008.

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

Bouma, Gerlof. "Normalized (pointwise) mutual information in collocation." Edited by Christian Chiarcos , Richard Eckart de Castilho and Manfred Stede. *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009.* Tubingen: Gunter Narr Verlag, 2009. 31-40.

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

J.-D. Kim, T. Otha, T. Tateisi i J.-I. Tsuji. „GENIA corpus -- a semantically annotated corpus of bio-textmining.” *Bioinformatics*, 2003: 180-182.

K. Toutanova, K. Klein, C. Manning i Y. Singer. „Feature-rich part-of-speech tagging with a cyclic dependency network.” *Proceedings of HLT-NAACL 2003.* 2003. 223-259.

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

Korkontzelos, Ioannis, Ioannis P. Klapaftis, and Suresh Manandhar. "Reviewing and evaluating automatic term recognition techniques." *Advances in Natural Language Processing.* Springer, 2008. 248-259.

Lossio-Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. "Yet Another Ranking Function for Automatic Multiword Term Extraction." *PolTAL 2014.* Springer, 2014. 52-64.

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

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

Marciniak, Małgorzata, and Agnieszka Mykowiecka. "Terminology extraction from medical texts in Polish." *Journal of Biomedical Semantics*, 5 30, 2014.

—. "Terminology extraction from domain texts in Polish." Edited by R Bembenik, L Skonieczny, Henryk Rybiński, M Kryszkiewicz and M Niezgódka. *Intelligent Tools for Building a Scientific Information Platform. Advanced Architectures and Solutions.* Springer, 2013. 171-185.

—. "Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish." *Proceedings of BioNLP 2011*. 2011. 92-100.

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

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

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

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

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

Wermter, Joachim, and Udo Hahn. "Massive biomedical term discovery." *Discovery Science.* Springer, 2005. 281-293.

Y. Tateisi i J.-I. Tsujii. „Part-of-speech annotation of biology reseach abstracts.” *Proceedings of 4th International Conference on Language Resources and Evaluation.* Lisbon, Portugal, 2004. 1267-1270.