Terminology Extraction from Domain Texts in Polish

Małgorzata Marciniak and Agnieszka Mykowiecka

Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
{mm,agn}@ipipan.waw.pl

Abstract. The paper presents a method of extracting terminology from Polish texts which consists of two steps. The first one identifies candidates for terms, and is supported by linguistic knowledge—a shallow grammar used for extracted phrases is given. The second step is based on statistics, consisting in ranking and filtering candidates for domain terms with the help of a C-value method, and phrases extracted from general Polish texts. The presented approach is sensitive to finding terminology also expressed as subphrases. We applied the method to economics texts, and describe the results of the experiment. The paper closes with an evaluation and a discussion of the results.

Keywords: terminology extraction, shallow text processing, domain corpora.

1 Introduction

Finding information in large text collections efficiently is tightly connected with the problem of formulating appropriate questions. Without knowing the specific terminology used in a particular domain it can be hard to localize the documents which contain the information needed. It is particularly true if words which are used in a query are ambiguous or if the specific subject we are looking for is rarely described. Availability of appropriate vocabularies can improve both the process of precise question formulation and proper document indexing, but multiplicity of different domains and their quick changes in time make the task of manual preparation of such resources practically infeasible. The solution to this problem can lie in automation of the dictionary creation process. In our paper we present the complete procedure of terminology extraction from specialized Polish texts and results we obtained for the chosen domain—economics.

Terminology extraction is a process of identifying domain specific terms from texts and usually consists of two steps. The first one identifies candidates for terms and is usually supported by linguistic knowledge. The second step, based on statistics, consists in ranking and filtering candidates for domain terms. An overview of existing approaches to automatic terminology extraction is included in [9]. The first step of the process is crucial in the sense that the result can contain only those phrases which are defined at this stage. In practice, from an

R. Bembenik et al. (Eds.): *Intell. Tools for Building a Scientific Information*, SCI 467, pp. 171–185.
DOI: 10.1007/978-3-642-35647-6_13

algorithmic point of view, the existing solutions differ at this stage only in the degree in which they use linguistic knowledge and in choice of the type of phrases they operate on. Minimalistic solutions totally neglect linguistic information [16] while others use POS tagging and simple shallow syntactic grammars [4]. Much more diversity is observed at the second processing stage when the initially identified phrases are filtered and ranked.

Although research on automatic terminology extraction has been carried on for many years now, very little was done in this area for Polish. The inflectional character of the language, together with the lack of specialized tools and resources, make processing Polish domain specific texts challenging. Till recently only the related task of collocation recognition has been explored, e.g [3]. In this paper we present an attempt to use slightly modified methods of automatic term extraction presented in [4] to unstructured Polish texts. In comparison to the first experiment which was performed on Polish clinical data [8], the texts which were analyzed in this experiment are better edited and are more similar in both sturucture and vocabulary to general newspaper texts.

In the terminology extraction process presented in this paper we adopt a corpus based approach. Our rough definition of a term is as follows: "a term is a nominal phrase which is used in the domain texts frequently enough to make it plausible that it represents something important which can be asked for by Internet readers and it does not occur equally frequently in texts on different subjects". We do not try to interpret these terms in relation to any external formal domain related knowledge, as we concentrate on collecting language constructs which can be identified within real texts. The normalization of these terms which may result in representing complex phrases as instances of relations between different objects, should constitute the subsequent processing level.

Our approach to terminology extraction consists of two standard steps enumerated above. At the first step we identify both single-word and multi-word nominal phrases using a simple shallow grammar operating on morphologically annotated text. Then, we use a slightly modified version of a popular approach of [4] which also allows us to recognize phrases occurring only inside larger nominal phrases. In this method, all phrases (external and internal) are assigned a C-value which is computed on the basis of the number of their occurrences within the text, the context in which they occur and their length. The procedure was performed separately on the selected set of specialized texts and on the texts which are representative for general usage of Polish (the balanced one million word subcorpus of NKJP [13]). As domain specific phrases, we chose those which are relatively more frequent in the first set than in the second one.

The paper is organized as follows. First we present a characteristic of the type of terminology we are interested in. In the section 3 problems connected with processing specialized texts are presented, together with the accepted solutions. Section 3 is devoted to the description of the types of phrases which are considered as candidates for terms while the next part of the paper presents the method of establishing a ranked list of all identified phrases. Section 6 contains the description of the obtained results.

2 Domain Term Characteristics

The first step towards terminology extraction is to define the exact goal, i.e. to define the type of phrases which can constitute domain related terms. But even before doing that, one has to decide what is to be considered as a domain term and what is not. The decision is far from being straightforward, as it can be easily seen while inspecting various already existing terminological resources. In standard dictionaries we can find general, not very precise definitions of a term and terminology. Terminology is defined for example as "a system of words used to name things in a particular discipline" or "the technical or special terms used in a business, art, science, or special subject" (http://www.merriam-webster.com/dictionary/), while a term is "a word or expression that is used in a particular variety of English or in realtion to a particular subject" [15], "a name, expression, or word used for some particular thing, especially in a specialized field of knowledge" (http://www.collinsdictionary.com/dictionary/), or "word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject" (http://www.merriam-webster.com/dictionary/term). In manually created terminology resources, the decision what to include in the terminology lexicon depends on the authors' judgment supported by a set of adopted rules. In computational approaches to terminology extractions, terms are usually nominal phrases which have the specified structure and which fulfill some defined selection criteria. The final evaluation of obtained results is again done by domain specialists either directly or indirectly by comparing them to already existing resources. In the result, what is considered a domain term and what is not, depends on human judgment of the importance of a particular phrase for the chosen domain and on its exact definition, as well as the accepted description level.

Using already defined terminology resources is not easy as frequently they do not contain the exact terms which are needed in the specific task. One of the problems is the fact that while in specialized dictionaries there are a lot of very detailed terms, some general expressions can still be omitted. For example, one of the phrases which is typically connected with medical care is in Polish *badanie USG*, frequently shortened to *USG*. The Polish equivalent of the MESH thesaurus (http://www.ncbi.nlm.nih.gov/mesh) —`http://slownik.mesh.pl`— containing 24359 general and 546931 compound descriptors, lists 154 terms in which the word *badanie* 'examination' occurs but there is no term *badanie USG*. Only one subtype of this kind of an examination is mentioned: *Badanie USG prenatalne* 'Ultrasonic Prenatal Diagnosis'. General information is placed under the heading *U*SG. Among those mentioned above 154 terms containing *badanie*, a generally used phrase *badanie diagnostyczne* 'diagnostic examination' is not listed at all.

What is common to all domain vocabularies is that the vast majority (if not all) of terms are noun phrases. Thus, we also decided to concentrate on these kinds of phrases, although there are approaches, like [14], in which verbal phrases are also taken into account.

The internal structure of terminological noun phrases can vary, but the number of construction types are limited. In Polish, domain terms most frequently have one of the following syntactic structures:

- a single noun or an acronym, e.g. *grawitacja* 'gravity', *PKB* 'GDP';
- a noun followed (or, more rarely, preceded) by an adjective, e.g. $stosunki_{n}$ $gospodarcze_{adj}$ 'economic relations' or $nowe_{adj}$ $technologie_{n}$ 'new technologies';
- a sequence of a noun and another noun in genitive: $prawo_{n,nom}$ $ciążenia_{n,gen}$ '(Reilly's) law of (retail) gravitation';
- a combination of the last two structures: $europejski_{adj}$ $rynek_{n,nom}$ $usług_{n,gen}$ $finansowych_{adj}$ 'European market for financial services';
- a noun phrase modified by a prepositional phrase, e.g. *wierzytelność podatnika wobec skarbu państwa* 'taxpayer liability to the State (Treasury)'.

Other features of Polish nominal phrases which complicate the recognition process are: relatively free word order, genitive phrase nesting: the sequences of genitive modifiers can have more than two elements, internal prepositional phrases, and coordination. In section 4 of the paper we present a shallow grammar defining noun phrase types which we consider to cover most Polish nominal domain related terms.

3 Linguistic Analysis of Domain Specific Texts

The process of automatic domain terminology extraction starts with gathering an appropriate corpora containing domain texts. The extraction procedure begins with text segmentation into tokens: words, numbers and punctuation marks. The next step is morphological analysis and disambiguation, which results in tagging all word forms with information about their base form, part of speech, and morphological feature values. For this purpose we can use publicly available Polish taggers: TaKIPI tagger [10] or Pantera [1]. They cooperate with different versions of the morphological analyser Morfeusz [17] (former SIAT and current SGJP respectively).

The quality of results obtained from a terminology extraction system strongly depends on the quality of the domain corpus, especially on the quality of its tagging. Texts which deal with specific domains usually contain a lot of words which are not present in general dictionaries, there are also specific token types and domain related acronyms or abbreviations. Processing them with general language tools trained on newspaper or literature data usually leads to incorrect descriptions of many tokens, so to obtain reliable results we have two choices: to train a tagger (a program which disambiguates morphological descriptions) for domain texts which requires a lot of correctly tagged data, or to correct data obtained with an available tagger. As for the first solution, the appropriate training data is currently unavailable, we had to choose the latter option.

In Morfeusz, an abbreviation is assigned a *brev* POS (part of speech) with characteristic if the full stop is necessary after it (attribute *pun* and *npun*).

We decided to extended the description of an abbreviation with information about the type of word or phrase it abbreviates. The type of an abbreviated phrase/word is necessary if we want to construct a grammatical phrase containing the abbreviation. So, we add an attribute *btype* that has following values: *nphr, nw, prepphr, adjw, etc.* that informs if the abbreviated is a word or a phrase (ending *w* or *phr* respectively) and gives a grammatical type of an abbreviated phrase or POS of the abbreviated word. For example *nr* (*numer* 'number') has attribute btype *nw*, while *SA* (*Spółka Akcyjna* 'joint-stock company') is of *nphr* type.

Taggers can only annotate tokens for which descriptions are available to them. To deal with out-of-dictionary words, both taggers can cooperate with the *Guesser* module [11] which suggests tags for words which are not analyzed by Morfeusz. As in the experiment described in [6] only 20% of suggested descriptions were fully correct (this means that: base form, POS and its complete morphological characterization are correctly assigned), we decided not to use this module. In our case, unknown tokens are assigned the *ign* tag. The Pantera tagger allows us to define a separate dictionary in which we can describe unknown tokens.[1] Strings defined in that dictionary cannot contain spaces nor punctuation marks such as periods. Thus the common abbreviation *m.in.* (*między innymi* 'among others') cannot be introduced into this dictionary. However it is possible to introduce the full declination of the foreign word *outsoursing* 'outsourcing' that is used with Polish endings: $outsoursingiem_{inst}$ or $outsoursingu_{gen}$.

Another method of improving tagger results, is to define rules in Spejd [12]. The advantage of this method is the possibility of taking into account contexts. For example in the following string *Dz.U.* that abbreviates the phrase *Dziennik Ustaw* 'Journal of Law' the string *U* is interpreted as the preposition 'at', but in this context it is the abbreviation of the word *ustawa* 'law'. The places where the description of *U* ought to be changed is recognized by the Spejd rule (1)[2] which is named *DZU*. The rule indicates the modified element after the *Match* string: the orthographic form of the string equal to *U*. The contexts are described after the keywords *Left* and *Right*. In this case only the left context is defined. It consists of two tokens: the first has the orographic form *Dz* and the second is the full stop (without spaces between tokens—indicated by *ns*). The new description of *U* is given after the *Eval* keyword, e.g.: the new base form *ustawa* 'law' and the new tag that describes the abbreviation of a noun word.

(1) Rule "DZU"
Left: [orth~"Dz"] ns [orth~"."] ns;
Match: [orth~"U"];
Eval: word(Modif-morf, "base:ustawa#ctag:brev:pun:nw#");

Spejd rules are particularly helpful in correcting some regular tagging errors in often occurring phrases. For example, in the frequent phrase *osobami zarządza-*

[1] TaKIPI tagger does not allow to use an external dictionary. Some methods of correcting TaKIPI results are described in [6].

[2] Spejd also provides different methods of correcting word annotations.

jącymi lub nadzorującymi emitenta 'persons managing or supervising the issuer' the Pantera tagger assigned the gender of participle *nadzorującymi* 'supervising' to impersonal masculine instead of feminine. This can be corrected by the rule (2), that changes the description of the morphological features of the matched element, including the gender. The same error occurs in phrases where instead of the word *emitenta* 'issuer$_{gen}$' there is another phrase like *(między) osobami zarządzającymi lub nadzorującymi a Spółką* '(between) persons managing or supervising and the company', so the rule (2) is applied independently to a right context.

(2) Rule "nadzorującymi"
Left: [orth~"osobami"] [orth~"zarządzającymi"] [orth~"lub"];
Match: [orth~"nadzorującymi"];
Eval: word(Modif-morf, "ctag:pact:pl:inst:m3:imperf:aff#");

4 Phrase Selection

For recognizing the selected types of nominal phrases, we defined a cascade of shallow grammars consisting of six small sets of rules being regular expressions in which morphological information is used.

The first set of rules describe the basic types of phrase elements. Head elements of nominal phrases can have tags denoting nouns, gerunds or nominal abbreviations (*subst, ger, brev:xx:nw, brev:xx:nphr*). The first two types are recognized as inflecting elements, so they should agree with adjectival modifiers in number, case and gender, while the other two do not inflect, so they can be combined with modifiers of different forms. Adjectival modifies are also divided into two classes. The first class describes those elements which show inflection, i.e. adjectives, past participles and a special kind of complex adjectives which are built up from a special adjectival form ending with '-o', a hyphen, and an ordinary adjective (e.g. *społeczno-ekonomiczny* 'socioeconomic'). The second, non-inflecting group consists of adjectival abbreviations, e.g. *ang. (brev:pun:adjw)* 'English'.

The second set of rules describe adverbial modifications of adjectives, while the third one describes adjectival phrases which can consist of up to five adjectives optionally separated by commas. The last adjective in a sequence can be preceded by a conjunction *i* 'and'. At the next level, nominal phrases which consist of a nominal element and an optional pre or post adjectival modification are formed. A nominal phrase has gender, number and case assigned on the basis of the characteristic of its main element. In the case when the modified element is of a non-inflecting type, the inflectional description of the phrase is assigned on the basis of the adjectival features.

The next level of complex phrase formulation consists in building sequences of nominal genitive modifiers. The possibility of placing adjective modifications after the genitive one is also described. The last grammar level accounts for nominal modification in nominative case (apposition) and for modification by prepositional phrases. The examples of the nominal phrases recognized by the rules are the following:

- cenę$_{n,acc}$ ropy$_{n,gen}$ 'price of oil';
- współczesna$_{n,nom}$ struktura$_{n,nom}$ systemu$_{n,gen}$ transportowego$_{n,gen}$ 'contemporary structure of the transportation system';
- poziom$_{n,nom}$ wykorzystania$_{n,gen}$ standardowych$_{adj,gen}$ jednostek$_{n,gen}$ rozliczeniowych$_{adj,gen}$ 'degree of utilization of standard units of account';
- cena$_{n,nom}$ na$_{prep}$ nowy$_{adj,acc}$ produkt$_{n,acc}$ 'price on a new product';
- cena$_{n,nom}$ równowagi$_{n,gen}$ kształtowana$_{ppas,nom}$ przez$_{prep}$ relację$_{n,acc}$ podaży$_{n,gen}$ 'price of equilibrium shaped relative to supply'.

Applying these general rules to our data resulted in a set of phrases which included an easily distinguishable subset of non-domain terms. These were phrases beginning with modifiers describing that a concept represented by a subsequent subphrase is occurring, desired or expected, for example, *(w) trakcie*$_n$ *sesji* 'during the session'. To eliminate such phrases we defined a set of words which were to be ignored during phrase construction and modified the set of rules accordingly. The excluded words belong to the following classes:

- general time or duration specification, e.g. *czas 'time', miesiąc 'month'*;
- names of months, weekdays;
- introductory/intension specific words, e.g. *kierunek 'direction', cel 'goal'*;
- general adjectives which can modify nearly every phrase, e.g. *inny 'other', sam 'alone', niektóry 'some', który 'that', każdy 'every', taki 'such'*.

The set of phrases obtained using this modified grammar constituted a starting point for terminology selection procedure. As Polish is an inflectional language, phrases which are identified within the text are of different forms (e.g. cenę$_{n,acc}$ ropy$_{n,gen}$, cena$_{n,nom}$ ropy$_{n,gen}$) so the usual processing stages like counting phrase frequencies and preparing a list of phrase types became difficult. To overcome this problem we produce an artificial base form of every identified phrase occurrence taking base forms assigned by the tagger, i.e. cena$_{n,nom}$ ropa$_{n,nom}$.

To allow for the recognition of terms which are nested inside other more complex terms, we add information about internal phrase structure, i.e. we mark limits of substrings matched by rules applied at the subsequent levels of the grammar cascade. The annotation style is minimalistic, i.e. only the end of the phrase and its type is marked by the '>' sign with the type name. A phrase can not be divided on a $>_a$ (adjective) marker. For the selected examples, the grammar output looks as follows:

- cena $>_n$ ropa $>_n$
  price $>_n$ oil $>_n$
- współczesny $>_a$ struktura $>_n >_t$ system $>_n$ transportowy $>_a >_t >_{ng}$
  contemporary $>_a$ structure $>_n >_t$ system $>_n$ transport $>_a >_t >_{ng}$
- poziom $>_n$ wykorzystać $>_n$ standardowy $>_a$ jednostka $>_n >_t$ rozliczeniowy $>_a >_n >_{ng}$
  degree $>_n$ utilization $>_n$ standard $>_a$ unit $>_n >_t$ accounting $>_a >_n >_{ng}$
- likwidacja $>_n$ zagraniczny $>_a$ konto $>_n >_t$
  cancellation $>_n$ foreign $>_a$ account $>_n >_t$.

On the basis of the structural information we can identify nominal subphrases. For example, in the second phrase enumerated above there are four such subphrases (given here in a properly lematized form): *współczesna struktura, struktura systemu, współczesna struktura systemu*, and *struktura systemu transportowego* 'contemporary structure', 'structure of the system', 'contemporary structure of the system', 'structure of the transportation system'. In the last example only one of two substrings is a proper subphrase: *zagraniczne konto* 'foreign account'.

5 Term Identification

The set of phrases constitute input data for the term selection algorithm. This stage aims to identify subterms which occur inside other terms—internal terms—and to eliminate phrases which come from general language and should not be placed within the domain dictionary. To eliminate phrases from general language, we compare frequencies of selected phrases in a domain corpus and in a corpus of general language.

For the purpose of ranking terms we adopted one of the most popular solutions to this problem proposed by [4], [2]. In this approach, all phrases (external and internal) are assigned a C-value which is computed on the basis of the number of their occurrences within the text and their length. Internal phrases are not just every substring of the identified phrases, but only those sequences of phrase elements which would be accepted by our grammar as correct nominal phrases (the exact identification process was characterized in the previous section). As we also wanted to take into account phrases of the length 1, for one word phrases we replace the logarithm of the length with the constant 0.1.[3] This slightly modified definition of the C-value is given below (p – is a phrase under consideration, LP is a set of phrases containing p):

$$C - value(p) = \begin{cases} lc(p) * freq(p) - \frac{1}{\|LP\|} \sum freq(lp), & if\ \|LP\| > 0,\ lp \in LP \\ lc(p) * freq(p), & if\ \|LP\| = 0 \end{cases}$$

$$\text{where } lc(p) = log_2(length(p)) \text{ if } length(p) > 1 \text{ and } 0.1 \text{ otherwise.}$$

The general idea of this coefficient is to promote phrases which occur in different contexts as it is more likely that they constitute separate terms than in the case where most of their occurrences have the same context. For example, *system bankowy* 'banking system' has occurred in the analyzed texts 5 times of which 5 were inside a wider nominal phrase of 5 different types, for *bezpieczeństwo publiczne* 'public security' these numbers were respectively 4-1-1. In both these cases we have clear evidence that these phrases constitute separate terms. Similarly, the phrase *waluta narodowy* 'national currency' which never occurred in

[3] The value 0.1 was chosen arbitrary from the interval 0–1 to balance lower frequencies of phrases of length 2 in comparison to single words. If many very long phrases are recognized within a particular data, the appropriate coefficient should also be modified as terms consisting of very many words practically do not occur.

isolation, but occurred five times in four different contexts, should be considered as a phrase. The phrase *wysoki przedział* 'high interval' would be much lower on the term ranking list as it occurred 3 times but only in one type of context.

The choice of the C-value coefficient was supported by evaluation done in [5] and [9] which showed its high efficiency for term identification task. Apart from the relatively high usage of this coefficient, there are nevertheless some problems in interpreting the notion of the LP set from the definition given above. If we consider a following set of phrases: *kapitał zakładowy*, 'share capital' *pozyskanie kapitału*, 'acquisition of capital' and *pozyskanie kapitału zakładowego* 'acquisition of share capital' it is not straightforward how to count contexts in which the basic phrase 'capital' occurs. According to the original definition of C-value method three different contexts would be counted. In our approach, to cover possible change of word order, we decided to count right and left contexts separately so we count only two.

6 Experiment Description

The experiment aimed at automatic extraction of domain specific terms was conducted on the economics articles taken from the Polish Wikipedia. Only textual content of these articles was taken into account. Texts were collected within the Nekst project (*An adaptive system to support problem-solving on the basis of document collections in the Internet* POIG.01.01.02-14-013/10), by Łukasz Kobyliński in 2011 in order to test word sense disambiguation methods. The data contains 1219 articles that have economics related headings and articles linked to them. The data contains about 450,000 tokens.

The plain texts were processed by Pantera tager working with Morfeusz SGJP. We defined, through Pantera, an additional domain dictionary containing 741 entries of word-forms (not recognized by the Morfeusz analyzer) and their descriptions. Additionally 156 Spejd rules were created to correct Pantera decisions, and to extend descriptions of abbreviations. Spejd rules corrected or extended about 5500 token descriptions.

As the reference set we have chosen the balanced subcorpus of NKJP [13]—*nkjp-e*. It was originally tagged using Pantera and then manually corrected, so for this set we did not prepared any additional dictionaries nor correction rules. The entire set consists of about 1,200,000 tokens.

The results of applying the grammar described in section 4 to the economics texts (*wiki-econo*) and to the general corpus texts (*nkjp-e*) are presented in Table 1 in which the distribution of phrase lengths and frequencies are given.

The list of phrases obtained after processing economics texts had to be cleaned up from two kinds of expressions. First, some phrases occurring within these texts are coming form general language and should not be treated as economic. The examples of these phrases could be: *pierwsza próba* 'the first trial', *sposób liczenia* 'the way of counting', and *czynnik zewnęrzny* 'external factor'. The second group of phrases are those which resulted from extracting internal nominal phrases and in practice never occur alone, for example *konwersja części* 'conversion of

Table 1. Distribution of phrases lengths and frequencies

phrase	data set		common		phrase	data set	
length	wiki-econo	nkjp-e	nb	%	freq	wiki-econo	nkjp-e
$\sum$	104847	232099	13359	12.74	$\sum$	104847	232100
1	8214	35249	6270	76.38	=1	85747	197713
2	39607	96088	6051	15.27	2-10	16623	29599
3	27826	53680	832	2.99	11-50	1885	3602
4	15980	27518	149	0.93	51-100	280	631
5	7836	12200	42	0.53	101-1000	305	545
6-9	5292	7257	15	0.28	1000-	7	10
>=10	92	107	0	0			
max	13	14	-	-	max	1565	2414

a part' being a part of *konwersja części długu* 'conversion of a part of debt' or *dokument wystawiony* 'documment issued' extracted from *dokument wystawiony przez podmiot* 'document issued by a given subject'.

The first problem is addressed by comparing the resulting data set to the list of phrases obtained for the general texts, while the second one is (partially) solved by ordering the phrases according to their C-value and eliminating those which are low on this list.

To compare the terminology extracted from economics texts with phrases extracted from the general corpus of Polish we analyzed terms identified in both corpora. Table 3 shows how many terms are recognized in both corpora and how many of them have greater C-value in each data set. 4% multi-word terms recognized in economics texts are also recognized in *nkjp-e* data—the longest common phrases have 6 words. 2.2% of multi-word terms have higher C-value in economic than in *nkjp-e* data. For economics texts C-value is higher for example for the phrase *papiery wartościowe dopuszczone do publicznego obrotu* 'securities admitted to public trading' while for *nkjp-e* subcorpus such a phrase is *minister właściwy do spraw finansów publicznych* 'minister responsible for public finances'. This example illustrates the observation that some of multi-word terms with higher C-value in *nkjp-e* data are in fact related to the economic domain. There are quite a lot of phrases with greater C-value for *nkjp-e* subcorpus and relevant to the economic domain, below a few more examples of such phrases are given:

- *walne zgromadzenie akcjonariuszy* 'general meeting of shareholders';
- *Narodowy Bank Polski* 'Polish National Bank';
- *narodowy fundusz inwestycyjny* 'national investment fund';
- *rada nadzorcza* 'supervisory board';
- *skarb państwa* 'state treasury';
- *urząd skarbowy* 'treasury office'.

In our opinion the situation described above results from the popularity of economic topics in newspaper articles and Parliamentary speeches which are included in the *nkjp-e* subcorpus. Because of that, it would be desirable to inspect

Table 2. The most frequent phrases

	wiki-econo			nkjp-e	
	phrase	occur.		phrase	occur.
1	cena 'price'	1565	1	pan 'mister'	2414
2	spółka 'company'	1364	2	człowiek 'human'	1789
3	rynek 'market'	1300	3	sprawa 'case'	1500
4	koszt 'cost'	1277	4	praca 'work'	1373
5	podatek 'tax'	1214	5	osoba 'person'	1196
...			...		
92	papier wartościowy 'stock'	281	386	pan poseł 'member of Parliament'	134
130	działalność gospodarcza 'economic activity'	220	556	unia europejska 'European Union'	100
192	osoba fizyczna 'natural person'	156	612	projekt ustawy 'project of an Act '	91
209	podatek dochodowy 'income tax'	148	641	minister właściwy 'appropriate minister'	87
244	fundusz inwestycyjny 'investment fund'	128	723	rada ministrów 'Council of Ministers'	79
...			...		
431	kodeks spółki handlowej 'code of commercial companies'	70	841	minister właściwy do spraw 'minister responsible for the task'	70
538	spółka z ograniczoną odpowiedzialnością 'limited liability company'	38	1600	Jan Paweł II 'John Paul II'	37
539	koszt uzyskania przychodów 'cost of revenues'	38	1739	II wojna światowa 'II World War'	33
540	prowadzenie działalności gospodarczej 'running a business'	38	1902	sojusz lewicy demokratycznej 'Democratic Left Alliance '	30
563	ustawa o rachunkowości 'Accounting Act'	33	2162	wejść w życie 'come into force'	26

Table 3. Comparison with general corpus

Terms	common	C-value greater in econom.	C-value greater in *nkjp-e*
1-word	5535	1563	3972
2-words	3526	1963	1563
3-6-words	360	224	136
Total	9421	3750	5671

manually phrases recommended for removing from the domain terminology on the basis of comparison with phrases created from *nkjp-e* subcorpus. However, as manual inspection of so many phrases is hard to perform, we decided to stick to automatic approach and eliminate from the result all phrases which have greater

Table 4. C-value distribution

C-value	initial nb of phrase types	nb after selection
=0	53513	51888
<1	4597	3357
<2	25239	24512
<5	18987	18551
<10	1454	1324
<100	1043	943
>=100	14	14
Total	104847	100589

Table 5. Phrases considered as terms

	1st annotator			2nd annotator			final annotation		
	domain	general	wrong	domain	general	wrong	domain	general	wrong
top	412	72	16	413	72	15	409	75	16
middle	322	135	43	290	141	69	278	263	59
end	246	170	84	209	181	110	206	187	107

C-value counted in the context of general texts than that counted for the economic data. After this step the term list consisted of 100589 nominal phrases with the distribution of C-value given in Table 4.

To evaluate the quality of the results we performed a manual verification of three groups of 500 randomly chosen phrases. The first set was drawn from the top terms, which have C-value greater or equal 5, while two others represent terms which have C-value below or equal 2 and above 1 (a set named *middle*) or equal 0 (*end*). These lists were checked by two annotators. They had to qualify each term either as a domain specific, general or wrong (i.e. sequences which are not terms at all or have wrong syntactic structure). The results of this verification are given in Table 5. In this table we can observe that the task of judging what is and what is not a domain terminology is highly subjective—the number of phrases judged as general differ a lot. Two examples of such differently judged phrases are *konkurs ograniczony* 'limited competition' and *matematyka stosowana* 'applied mathematics'. But in spite of these differences, the obtained results show that the applied method of automatic term extraction can give reliable results. More than 80% of terms from the *top* list are judged as domain related by both annotators, while this percentage lowers significantly towards the end of the ranked list. 40% of domain related terms in the *end* group of phrases is probably the result of the rather small size of a data set.

To confront automatic extraction with manual dictionary creation, we compared the results of our experiment with the economic terms dictionary constructed by Agata Savary within the already mentioned Nekst project. It contains terminology manually collected from different economics dictionaries and consists of about 10,000 multi-word terms. All possible grammatical forms of terms were created with help of *Toposław* tool [7].

To perform the comparison, from our list of terms that have C-value grater than in NKJP data, we removed terms that have C-value less then 1. Because the manually constructed dictionary consists only of multi-word phrases, we deleted also one word terms. To perform the comparison of such prepared list of terms with the manually created dictionary we had to identify all their occurrences in texts. Then, we compared our list of terms (their occurrences) with that manually collected and declined. The results are given in Table 6. As we can see about 90% phrases recognized in manually created dictionary was represented in our texts as full phrases while 10% appeared only as subphrases.

Table 6. Comparison with manual dictionary

Phrases	our method	in manual. dict.
full phrases	41031	2142
subphrases	2943	245
Total	43974	2387

The results of this comparison show that on our list there are many terms which should be taken into account in economics dictionary. For example the manually created dictionary contains 25 different phrases describing *aktywa* 'assets' (we considered phrases where 'asset' is the head element), 7 of them were recognized in our data: *aktywa finansowe* 'financial assets', *krótkotrwałe aktywa finansowe* 'short-term financial assets', *aktywa obrotowe* 'current assets' and *aktywa trwałe* 'fixed assets', *aktywa netto* 'net assets', *oficjalne aktywa rzeczowe* 'official tangible assets', *aktywa rezerwowe* 'reserve assets' . Another 25 phrases describing assets (not present in the dictionary) were recognized by our method. 20 phrases are correct domain terms like: *aktywa firmy* 'assets of the company', *zagraniczne aktywa* 'foreign assets', *aktywa niefinansowe* 'non-financial assets', *aktywa trwałej wartości* 'assets of lasting value', *płynne aktywa* 'liquid assets'. 2 phrases are not classified as the domain terms: *wszystkie aktywa* 'all assets' and *pozostałe aktywa* 'other assets' while 3 phrases are incorrect.

7 Conclusions

In this paper we presented results of automatic terminology extraction from domain unstructured texts. Such tools can be valuable for processing texts from domains for which no electronic terminological or ontological resources exist. Although good results presented in this paper were obtained on the basis of relatively clean data—grammar rules operate on the results of a general tagger which were corrected by a set of dedicated rules, the method can also be used on uncorrected data. In that case the results would be worse but they still may be of a practical value. The performed comparison with the manually created terminological lexicon showed that automatic terminology extraction can be also a valuable method for enriching already existing dictionaries, especially in domains which quickly change in time.

Although using the adopted method, a lot of relevant phrases can be identified, the selection procedure should be further improved. In further work we plan to enhance a definition of potential term structure and to define more sophisticated rules of candidates ranking which would be more suitable for Polish phrases.

Acknowledgments. This work was supported by SYNAT project financed by the Polish National Center for Research and Development (SP/I/1/77065/10).

References

1. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) IceTAL 2010. LNCS, vol. 6233, pp. 3–14. Springer, Heidelberg (2010)
2. Barrón-Cedeño, A., Sierra, G., Drouin, P., Ananiadou, S.: An Improved Automatic Term Recognition Method for Spanish. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 125–136. Springer, Heidelberg (2009)
3. Broda, B., Derwojedowa, M., Piasecki, M.: Recognition of structured collocations in an inflective language. System Science (4) (2008)
4. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. Journal on Digital Libraries 3, 115–130 (2000)
5. Korkontzelos, I., Klapaftis, I.P., Manandhar, S.: Reviewing and Evaluating Automatic Term Recognition Techniques. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 248–259. Springer, Heidelberg (2008)
6. Marciniak, M., Mykowiecka, A.: Towards morphologically annotated corpus of hospital discharge reports in Polish. In: Proc. of the BioNLP, ACL/HLT 2011 Workshop, Portland, Oregon (2011)
7. Marciniak, M., Savary, A., Sikora, P., Woliński, M.: Toposław – A Lexicographic Framework for Multi-word Units. In: Vetulani, Z. (ed.) LTC 2009. LNCS, vol. 6562, pp. 139–150. Springer, Heidelberg (2011)
8. Mykowiecka, A., Marciniak, M.: Terminology extraction from medical texts in Polish. In: Ananiadou, S., Pyysalo, S., Rebholz-Schuhmann, D., Rinaldi, F., Salakoski, T. (eds.) Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, SMBM 2012 (2012)
9. Pazienza, M.T., Marco Pennacchiotti, M., Zanzotto, F.M.: Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In: Sirmakessis, S. (ed.) Knowledge Mining. STUDFUZZ, vol. 185, pp. 255–279. Springer, Heidelberg (2005)
10. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly 11(1-2), 151–167 (2007)
11. Piasecki, M., Radziszewski, A.: Polish Morphological Guesser Based on a Statistical A Tergo Index. In: Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA 2007), pp. 247–256 (2007)
12. Przepiórkowski, A.: Powierzchniowe przetwarzanie języka polskiego. Akademicka Oficyna Wydawnicza EXIT, Warsaw (2008)
13. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw (2012)
14. Savova, G.K., Harris, M., Johnson, T., Pakhomov, S.V., Chute, C.G.: A data-driven approach for extracting “the most specific term" for ontology development. In: Proc. of AMIA (2003)

15. Sinclair, J. (ed.): Collins Cobuid English Language Dictionary. Collins Publ. (1990)
16. Wermter, J., Hahn, U.: Massive Biomedical Term Discovery. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS (LNAI), vol. 3735, pp. 281–293. Springer, Heidelberg (2005)
17. Woliński, M.: Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In: Kłopotek, M., Wierzchoń, S., Trojanowski, K. (eds.) Intelligent Information Processing and Web Mining, IIS: IIPWM 2006 Proceedings, pp. 503–512. Springer, Heidelberg (2006)