Multi-word Term Extraction for Bulgarian

Svetla Koeva
Department of Computational Linguistics – IBL
Bulgarian Academy of Sciences
52 Shipchenski prohod Blv. Sofia 1113, Bulgaria
svetla@ibl.bas.bg

Abstract

The goal of this paper is to compile a method for multi-word term extraction, taking into account both the linguistic properties of Bulgarian terms and their statistical rates. The method relies on the extraction of term candidates matching given syntactic patterns followed by statistical (by means of Log-likelihood ratio) and linguistically (by means of inflectional clustering) based filtering aimed at improving the coverage and the precision of multi-word term extraction.

1 Introduction

The goal of this paper is to compile a method for multi-word term extraction, taking into account both the linguistic properties of Bulgarian terms and their statistical rates. Term extraction exploits well-established techniques that seem difficult to improve significantly. As in many other areas of computational linguistics, term extraction has been approached generally with three different strategies – linguistic techniques, statistical techniques and a combination of both (Bourigault *et al.*, 2001; Jacquemin & Bourigault, 2000). The linguistically based techniques exploit the morpho-syntactic structure of terms that usually differ from one language to another (for example in Bulgarian and in English the most frequent syntactic structure representing terms is the noun phrase, but the two languages significantly differ in their constituent structure and agreement properties). The automatic extraction of term morpho-syntactic patterns, being in most cases language-dependent, requires specific language processing – Part-of-speech (POS) tagging, lemmatization, syntactic parsing, etc. The statistical techniques, on the other hand, rely on the different statistical features of terms compared to other words in the text and are usually based on the detection of words and expressions with a frequency value higher than a given limit. Some of the statistical approaches focus on the association measures between the components of the multi-word terms. Hybrid approaches, combining linguistic and statistical techniques, are also applied, mainly in two manners: statistical proceeding is used to filter the term candidates obtained through linguistic techniques, and, vice versa, some linguistic filters are exploited after statistical processing, in order to extract the statistically significant word combinations that match some given syntactic patterns.

The method for automatic multi-word term extraction, presented in this paper, also relies both on linguistic knowledge and on statistical processing. The research aims are to:

- Apply syntactic patterns of Bulgarian terms directed to multi-word term extraction;
- Use well-known statistical methods (association measures) to eliminate some of the irrelevant multi-word terms;
- Further limit the number of invalid terms by clustering term candidates around their lemmas;
- Test the performance of such a method over the manually annotated corpus.

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 59–66,
Prague, June 2007.

Most of the current methods for automatic term extraction are developed for English, and thus they are not appropriate for direct adaptation to Bulgarian, due to the morpho-syntactic differences between the two languages. Bulgarian is a language with a rich inflectional system. That is to say, a noun lemma can appear in six forms if it is masculine and in four forms if it is feminine or neuter. Besides, noun phrase structure and agreement properties in Bulgarian differ in some aspects from other languages such as English, Therefore, a language-specific approach is needed if we want to utilise the morpho-syntactical information for term extraction. To the best of our knowledge there is no report of an extensive work directed towards Bulgarian term extraction.

The structure of our paper outlines the three steps involved in our approach. In the following section we present a short linguistic analysis of Bulgarian terms. In the third section, we describe the identification of the candidate terms. The fourth section explains how we applied a list of terms to the filters. We then evaluate our results on a corpus that was set up by manual annotation. Finally, we discuss some peculiarities of the presented study and propose future works to be done.

2 Linguistic analysis of Bulgarian terms

2.1. Compilation of a term annotated corpus

We share the views that larger corpora not only give statistically more reliable counts, but also reveal phenomena that are completely lacking in smaller samples. The Acquis Communautaire (AC)[1] – the European Union legislation, which consists of approximately eight thousand documents containing approximately 40 million words (to be more specific, its Bulgarian subpart) – is targeted as the most appropriate resource for our research: because of its size, and because of the number of languages included in it. (The proposed method can be further transformed and/or evaluated to deal with the rest of the languages represented in the parallel corpus.)

The AC contains documents from several domains, which are divided into chapters: Agriculture, Fisheries, Transport Policy, Taxation, Economic and Monetary Union, Statistics, Social Policy and Employment, Energy, Industrial Policy, Education and Training, Telecommunication and Information Technologies, Culture and Audiovisual Policy, etc. This annotated subpart of the Bulgarian AC is developed as a test corpus and contains 10,521 words from randomly selected text samples representing the domains of Agriculture (AGR), Energy (ENR) and Education and Training (EDC).

Some criteria for the manual annotation of Bulgarian terms were defined, the notion of term among others. As with most linguistic concepts, a term is defined in various ways. For example, as "a word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession" (Webster, 2002), or as "a word or expression used for some particular thing"[2], or generally as words or phrases that denote specific concepts in a given subject domain. For the purposes of this investigation we defined a term as

An open class word or expression that is peculiar to a specific domain of human activities and occurs with a determinate (in some limits) frequency in that domain.

The annotation of terms in the Bulgarian AC subpart is also based on both the maximum and minimum length term selection. That is, in the case of a multi-word term which constituents are also terms, the longest term (as well as all shorter terms) is selected. It should be pointed out, however, that the term annotated corpus is still small enough to be representative of the word frequency and is a sample of translated texts that might manifest different tendencies for a term's distribution from those in the original texts.

2.2. Single-word terms vs. multi-word terms

The general impression is that the most of the papers dealing with automatic term extraction (especially the statistically based ones) are focused on multi-word terms. This can be explained by the fact that for English a bigger percentage of multi-word terms comparing to single-word terms is reported. To show the tendency for the correlation between single-word and multi-word terms in Bulgarian texts, the manually annotated subpart of the Bulgarian AC has been studied. We found out (Table 1.) that the proportion of single-word terms

[1] There has been some experience of exploiting the AC as a multilingual corpus (Steinberger at al., 2006).

[2] http://wordnet.princeton.edu

varies from about 2.5% to 3% depending on the subject domain.

The results show that the use of single-word terms in Bulgarian technical documents is also not very frequent and the tendency is that multi-word terms are preferred to single-word ones. Following these observations, first we will concentrate on the extraction of the Bulgarian multi-word terms.

Domain	AGR	ENR	EDC	Total
#Words	4423	3002	3096	10521
#Terms (T)	344	297	254	895
#Multi-word T	266	165	171	602
#Single-word T	111	89	93	293
% Terms	7,77	9,89	8,2	8,5
% Single-word T	2,5	2,96	3	2,78

Table 1. Distribution of single-word terms

2.3 Syntactic structures of Bulgarian terms

The starting point for the linguistically motivated part of the automatic term extraction is to describe the syntactic structure of Bulgarian terms. There are several Bulgarian terminological dictionaries published and some terminological databases available on the internet – all recourses are taken into consideration in the analysis without providing exact calculations. The collection of Bulgarian terms, obtained by the annotated subpart of the Bulgarian AC, is used as a source for the determination of the most frequent syntactic structures of Bulgarian terms.

It is claimed that NPs constitute about 80-99 % of whole terms in an English text, with the varying percentage depending on the text types (Arppe, 1995). The same statement is roughly true for Bulgarian; although there are some adjectives and verbs that can be regarded as terms in a certain domain (only three verbs and one adjective are detected in the annotated corpus). In this study we have concentrated on the NPs' term extraction, which comprises the focus of interest in several studies (Jacquemin, 2001; Justeson & Katz, 1995; Voutanen, 1993).

In order to obtain the statistics, the annotated part of Bulgarian AC is pre-processing. This allows the consequences of the categories constituting Bulgarian terms to be extracted and their frequency to be calculated. As a result, 16 different sequences of categories are obtained, among them 5 with a rate higher than 11 %. In the next examples the most frequent syntactic patterns of the Bulgarian multi-word terms are listed following their frequency rate:

- AN → *riboloven sezon* (fishing season), *iglolistno darvo* (conifer), *zemedelski ceni* (firm prices), *termalna energiya* (thermal energy), *klimatichna instalaciya* (air-conditioning);
- NpN → *obogatyavane na gorivo* (fuel enrichment), *podobryavane na pochvata* (soil improvement), *prava na deteto* (children's rights), *svoboda na pechata* (freedom of the press);
- NpAN → o*pazvane na okolnata sreda* (environmental protection), *nomenklatura na zemedelskite produkti* (agricultural product nomenclature), *izpolzvane na slanchevata energiya* (solar energy end-use applications), *sredstva za masova informaciya* (media);
- AAN → *semeyno zemedelsko stopanstwo* (family farming), *evropeyska parichna sistema* (European Monetary System), *inteligentna transportna sistema* (intelligent transport system), *magniten informacionen nositel* (magnetic medium);
- ANpN → *elektronen transfer na fondove* (electronic funds transfer), *opticheskо razpoznavane na simvoli* (Optical Character Recognition), *pravna uredba na telekomunikaciite* (regulation of telecommunications), *izbiratelno razprostranenie na informaciya* (selective dissemination of information).

Among the five types, the AN structure was the most frequent one, although the exact percentage still remains to be calculated over the bigger corpus.

The main differences observed concerning these five Bulgarian structures and their English equivalents are the regular agreement between the adjectival modifier and the head noun in Bulgarian and the prepositional phrase in Bulgarian instead the noun modifier in English. The adjective-noun agreement in Bulgarian noun phrases is partially exploited in the presented piece of work, but it might be extensively considered in further improvements of the method.

In the case of NpN, NpAN and ANpN structures, we found out that most of the terms corresponding to these patterns are built up with the Bulgarian preposition *na* (of). This may be explained by the fact that these PPs usually correspond to the English NPs with a noun modifier denoting more specific concepts. The possible strings of categories that might constitute the Bulgarian terms are exploited due to the fact that Bulgarian terms usually do not allow other constituents among their parts.

2.4 Term variations

Some authors have pointed out the discrepancy between term representation in dictionaries, and the term forms used in real texts (Daille, 2003). It is well known that the same concept can be formulated in different ways and the automatic term extraction should be able to recognize and link those different linguistic forms or expressions. Different kinds of term variants are distinguished in the literature: orthographic variants (capitalization), inflectional variants (word forms), morpho-syntactic variants (derivation), syntactic variants (word order differences) and semantic variants (synonyms).

In this study only the orthographic and inflectional variants are taken into consideration. It should be pointed out that compared to lemmas the multi-word terms have their own inflective rules. The POS of the head word determines the clustering of the term into grammatical classes, such as noun, adjective, and so on, which define the possible slots in the paradigm.

The significant grammatical categories inherent to the lemma of the head word (such as gender for nouns), the number and POS of the remaining constituents and the options for inserting some words (such as particles) in the multi-word term structure all show the grouping of multi-word terms' grammatical subclasses and define which slots of the paradigm are realized in the language. And finally, the formation of word forms of each component of a multi-word term and the type of agreement dependencies between components show the classification of multi-word terms into grammatical types that describe the real word paradigm belonging to a particular term (Koeva, 2005).

For instance, the Bulgarian term *klimatichna instalaciya* (air-conditioning) is a noun phrase; the members of the paradigm are determined by the head feminine noun. The inflection type is determined by the inflectional alternations of each member (the adjective and the noun):

klimatichna instalaciya – singular, indefinite
klimatichnata instalaciya – singular, definite
klimatichni instalaciii – plural, indefinite
klimatichnite instalaciii – plural, definite

There are agreement dependencies between adjective and head noun and no other words' intervention or word order changes are allowed.

3 Automatic term extraction

3.1 Pre-processing of the Bulgarian AC

It is common practice to extract candidate terms using a part-of-speech (POS) tagger and an automaton (a program extracting word sequences corresponding to predefined POS patterns). The part-of-speech tagging is the process of automatically identifying the words in a text as corresponding to a particular part of speech. The part-of-speech tagger used in this study is developed utilizing a large manually annotated corpus consisting of 197,000 tokens (150,000 words) randomly extracted from the Bulgarian Brawn corpus (1,000,000 words) (Koeva *et al.,* 2006). The tagger has been developed as a modified version of the Brill tagger (Brill, 1994). The Brill tagger was trained for Bulgarian using a part of the tagged corpus. We applied a rule-based approach leading to 98.3% precision. A sophisticated tokenizer that recognizes sentence boundaries and categorizes tokens as words, abbreviations, punctuation, numerical expressions, hours, dates and URLs has been built as a part of the tagger. For each word in the text the initial (most probable) part of speech among the ambiguity set is assigned from a large inflectional dictionary (Koeva, 1998).

The words that are not recognized by the dictionary are handled by the guesser analyzing the suffixes of the unrecognized words and assigning the initial part of speech among the ambiguity set. The part-of-speech ambiguity ratio calculated over the annotated corpus is 1.51 tags per word, which means that on average every second word is ambiguous. For solving the ambiguity, 144 contextual rules are implemented, utilizing the part of speech and dictionary information on the context, Some additional techniques for the optimizations are implemented – the application of dictionaries of abbreviations, proper nouns, grammatically unambiguous words, etc. After POS tagging the text remains unchanged and the additional information is added in an xml format.

Lemmatization is the process of automatic determining the lemma for a given word. Since the lemmatization involves fixing the part of speech of a word, it requires the running of a tagger. Lemmatization is closely related to stemming. The difference is that a stemmer operates on a single word without explicit knowledge of its identity as a part of speech, its lemma or its inflectional properties. For Bulgarian a large inflectional dictionary is used both for lemmatization and stemming.

The tag sets differ both in how the words are divided into categories, and in how their categories are defined. For the purposes of this investigation the grammatical information characterizing the forms is also assigned to nouns and adjectives, because the adjective-noun agreement is exploited.

3.2 Extraction of term candidates

Following the frequency analysis of the constituent structure of the Bulgarian multi-word terms, the targeted syntactic patterns will be recognized by the following regular expression:

[(A+N(pA*N)?)(NpA*N)]

The strings of categories bellow will be matched; those with more than two adjectives are either rare, or not observed in the language:

AN, AAN, NpN, NpAN, ANpN, ANpAN, NpAAN, ANpAAN, AANpAAN, …

The regular expression does not match the single Ns as well as the NPs with low frequently – only the five syntactic patterns with the highest frequency rate are targeted for the term extraction. Moreover, the agreement features of the Bulgarian NP structures are exploited considering the unification of grammatical features between the preceding adjective and the immediate following adjective or noun. Based on patterns' matching, the term candidates corresponding to the above regular expressions are extracted:

- AN → *osnovno obrazovanie* (basic education),
- AAN → *novi obrazovatelni metodi* (new educational methods), *evropeyska audiovizualna zona* (European audiovisual area),
- NpN → *ezik za programirane* (programming language),
- NpAN → *planirane na uchebnata godina* (planning of the school year), *elekronna obrabotka na danni* (electronic data processing), *potrebitel na ingormacionna tehnologiya* (information technology user), etc.

On the other hand, the following phrases (which are annotated as terms) are not recognized:

- NpVpN → *aparat za vazproizwodstvo na zvuk* (sound reproduction equipment),
- AcAN → *poshtenski i telekomunikacionni uslugi* (postal and telecommunications services),
- NpNpNN → *sistema za upravlenie na baza danni* (database management system), etc.

A deficiency of the approach based on the syntactic patterns is also the fact that any NP that matches the patterns will be selected as a term candidate, as is shown in the following examples:

- AN → *novi metodi* (new methods), *ogranicheno dvizhenie* (limited circulation),
- NpN → *analiz na informaciya* (information analysis), *broy na uchenicite* (number of pupils), etc.

Some of the noun phrases are wrongly extracted, although in this case this is concerned with a compositional building of structures that cannot be considered as that of multi-word terms. Some term candidates with a preposition cannot be treated even as phrases, because their parts belong to different sentence constituents. The identification of the sub-phrases that are themselves also terms should also be taken into account. In the following example, *sistema za upravlenie na baza ot danni* (database management system), the phrases *sistema za uprawlenie* (management system), *upravlenie na baza ot danni* (database management) and *baza ot danni* (database) are also terms.

Domain	AGR	ENG	EDC	Total
#Words	4,423	3,002	3,096	10,521
#Term candidates	901	778	712	2,391

Table 2. Number of term candidates

The number of extracted term candidates depends on the structure of the sentences that occur in the selected domains. Table 2 shows the extracted term candidates from a Bulgarian AC subpart representing texts from the Agriculture, Energy and Education domains.

4 Filtering of term candidates

As a filtering mechanism we adopted the calculating of the associativity between words, which is often used to identify word collocations, and the term clustering according to the inflexional paradigms.

4.1 Statistical filtering

The frequency-based techniques applied to term filtering assign a numerical value to sets of words to rank term candidates and exclude those term candidates below a certain threshold. The statement that the more frequently a lexical unit appears in a given document the more likely it is that this unit has a terminological function can be applied to certain genres of texts. Alone, frequency is not a robust metric for assessing the terminological property of a candidate.

In our case, we want to measure the cohesion of a multi-word candidate term by verifying if its words occur together as a coincidence or not. Association measures are often used to rate the correlation of word pairs (Daille, 1995; Daille *et al.*, 1998).

	B	!B	
A	N_{ii}	N_{ij}	N_{1p}
!A	N_{ji}	N_{jj}	N_{2p}
	N_{p1}	N_{p2}	N_{pp}

Table 3. The contingency table

These measures can be derived from the contingency table (Table 3.) of the word pair (A,B) containing the observed frequencies of (A,B), as follows:

N_{ii} = the joint frequency of word A and word B;

N_{ij} = the frequency word A occurs and word B does not;

N_{ji} = the frequency word B occurs and word A does not;

N_{jj} = the frequency word A and word B do not occur;

N_{pp} = the total number of ngrams;

N_{p1}, N_{p2}, N_{1p}, N_{2p} are the marginal counts.

The lexical association measures are formulas that relate the observed frequencies to the expected frequency (Mij = (Np1 * N1p) / Npp) under the assumption that A and B are independent. For the current work, the Log-likelihood coefficient has been employed (Dunning, 1993), as it is reported to perform well among other scoring methods (Daille, 1995).

$$\text{Log-likelihood} = 2 * \sum (N_{ij} * \log(N_{ij} / M_{ij}))$$

This calculation over the text serves as an important technique in identifying term candidates. The larger the value of Log-likelihood is, the stronger is the association between the two pairs of the string; consequently the string is the most probable candidate. Statistic filtering is applied only to those term candidates extracted by the linguistic component. For the calculation, the Ngram Statistics Package (NSP), programs that aids in analyzing ngrams, is executed (Banerjee & Pedersen, 2003). The NSP takes text files (in our case Cyrillic letters are transliterated into Latin) as input and generates a list of bigrams along with their frequencies as outputs. Over the list of bigrams obtained, the Log-likelihood is run to compute a ratio for each ngram. The bigrams are targeted because some of the term candidates initially extracted are long ones containing sub-phrases that are likely to function as term candidates. In order to avoid potential term candidates being included in other longer phrases, the term candidates are split and the constituting bigrams are generated.

As a result of statistical filtering, the initially selected term candidates are assigned different values according to their word association. The Log-likelihood coefficient computed for each bigram is used to decide whether or not there is enough evidence to reject or accept a bigram - there is a clear opposition between small and big values. Below the first five ranked candidates are listed.

1. *evropeyskata obshtnost* (European community)
2. *atomna energiya* (nuclear energy)
3. *detska gradina* (kindergarten)
4. *Darzhaven vestnik* (government newspaper)
5. *obrazovatelna sistema* (educational system)

4.2 Linguistic filtering

The linguistic filtering aims at linking the different variations of the same basic term. The list of the automatically extracted terms was reviewed by means of lemmatization in order to refine it and to increase the score of some terms. Until this stage the different word forms of a term were calculated separately. Bulgarian is a highly inflected language – the forms of the head noun can vary from one to seven depending of the gender, number and references to a person. The sequences of lemmas belonging to the term candidates are processed and the frequency values are recalculated according to the grouping of terms in one inflectional cluster with respect to the common canonical form. Through this technique morphologically-related occurrences, such as *iglolistno darvo* (a conifer), *iglolistnoto darvo* (the conifer), *iglolistni darveta* (conifers) and *iglolistnite darveta* (the conifers) are treated as one term.

5 Evaluation

The presented method of identifying Bulgarian multi-word terms was applied on the manually annotated corpus. First the texts were pre-processed by means of POS tagging and lemmatization, then the target syntactic patterns were extracted, and the rates of the related bigrams were calculated by means of Log-likelihood association, and finally additional reordering of term candidates was performed by means of inflectional clustering. As a result, 430 (from 539) correctly extracted multi-word terms are obtained – the precision of 79.96% is registered.

6 Conclusions and future work

We have presented a method aimed at extracting Bulgarian multi-word terms, which relies on the extraction of syntactic patterns from text and on the statistical and linguistically based filtering aimed at improving the coverage and the precision of multi-word collocation extraction. We have applied Log-likelihood ratio statistical filtering to the extracted multi-word terms. All extracted term candidates are grammatically correct, due to the syntactically based pattern matching. Further developments of the method include:

- Statistical determination of single-word terms;
- Coverage of long-distance occurrence and rare syntactic structures of multi-word terms;
- Analyzing the embedded terms.
- Using 'stop lists' of open and closed class words that are hardly to be found in the multi-word terms.

Some other experiments will be made using other well-known techniques of association measure. For the evaluation purposes the test corpus will be extended. A bigger homogeneous corpus would undoubtedly result in an increase in terms with more representative frequencies, and, therefore, in an improvement in statistical estimation of terms. The results can be exploited in the multilingual term extraction, due to the fact that the AC represents the biggest multilingual parallel corpus.

References

A. Aprre 1995. Term Extraction from Unrestricted Text: *10th Nordic Conference of Computational Linguistics (NoDaLiDa)*, Helsinki.

S. Banerjee and T. Pedersen 2003. The Design, Implementation, and Use of the Ngram Statistics Package, *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

D. Bourigault, C. Jacquemin, and M.-C. L'Homme 2001. *Recent Advances in Computational Terminology*, volume 2 of Natural Language Processing, John Benjamins.

E. Brill 1994. Some Advances In Rule-Based Part of Speech Tagging *AAAI,* Seattle, Washington

B. Daille 1995. *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Technical paper. UCREL, Lancaster University.

B. Daille 2003. Conceptual structuring through term variations, *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

B. Daille, E. Gaussier, and J.-M. Lange 1998. An Evaluation of Statistical Scores for Word Association, in J. Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds), *The Tbilisi Symposium on Logic, Language and Computation: Selected Papers*, CSLI Publications, p. 177-188.

T. Dunning 1993. Accurate methods for thestatistics of surprise and coincidence,. *Computational Linguistics*, 19(1):61–74.

C. Jacquemin 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

C. Jacquemin and D. Bouricault 2000. Chapter 19 Term Extraction and Automatic Indexing, *Handbook of Computational Linguistics* (R. Mitkov (ed.*)*), Oxford University Press, Oxford.

J. S. Justeson and S. M. Katz 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text, *Natural Language Engineering*. 1(1):9-27.

S. Koeva 1998. Bulgarian Grammatical dictionary. Organization of the language data, *Bulgarian language*, vol. 6: 49-58.

S. Koeva 2005. Inflection Morphology of Bulgarian Multiword Expressions, *Computer Applications in Slavic Studies – Proceedings of Azbuki@net, International Conference and Workshop*, Sofia, 201-216.

S. Koeva, S. Leseva, I. Stoyanova, E. Tarpomanova, and M. Todorova 2006. Bulgarian Tagged Corpora, *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages,* Sofia, 78-86.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC'2006), Genoa.

A. Voutilainen. 1993. NPtool. A detector of English noun phrases, *Proceedings of the Workshop on Very Large Corpora*, Columbus, Ohio.