Morpho-syntactic Clues for Terminological Processing in Serbian

Goran Nenadić
Department of Computing
UMIST, UK
G.Nenadic@umist.ac.uk

Irena Spasić
Computer Science
University of Salford, UK
I.Spasic@salford.ac.uk

Sophia Ananiadou
Computer Science
University of Salford, UK
S.Ananiadou@salford.ac.uk

Abstract

In this paper we discuss morpho-syntactic clues that can be used to facilitate terminological processing in Serbian. A method (called SRCE) for automatic extraction of multiword terms is presented. The approach incorporates a set of generic morpho-syntactic filters for recognition of term candidates, a method for conflation of morphological variants and a module for foreign word recognition. Morpho-syntactic filters describe general term formation patterns, and are implemented as generic regular expressions. The inner structure together with the agreements within term candidates are used as clues to discover the boundaries of nested terms. The results of the terminological processing of a textbook corpus in the domains of mathematics and computer science are presented.

1 Introduction

An overwhelming amount of textual information presented in newswire, scientific literature, legal texts, etc., makes it difficult for a human to efficiently localise the information of interest. In particular, it is doubtful that anybody could process such huge amount of information without an automated help, especially when the information content spans across domains. The amount of e-documents and their fuzzy structure require effective tools that can help users to systematically gather and make use of the information encoded in text documents. For these reasons, different text and/or literature mining techniques have been developed recently (e.g. (Hearst et al., 2000; Grobelnik et al., 2000)) in order to facilitate efficient discovery of knowledge contained in large scientific or legal text collections. The main goal is to retrieve the knowledge "buried" in a text and to present it to users in a digested form.

The discovery (and transfer) of knowledge relies heavily on the identification of relevant concepts, which are linguistically represented by domain specific terms. *Terms* represent the most important notions in a domain and characterise documents semantically, and thus should be used as a basis for sophisticated knowledge acquisition. Still, few text-mining systems incorporate deep and dynamic terminology processing, although there is an increasing amount of new terms that represent newly created concepts in rapidly developing fields. Existing term dictionaries and standardised terminologies offer only a partial solution, as they are almost never up-to-date. Although naming conventions do exist for some types of concepts (e.g. gene and protein names in biomedicine), these are only guidelines and as such do not impose restrictions to domain experts, who frequently introduce ad-hoc terms. Thus, the lack of clear naming conventions makes the automatic term recognition (ATR) task difficult even for languages that are not morphologically and derivationally rich.

ATR tools have been developed for English (Frantzi et al., 2000), French (Jacquemin, 2001), Japanese (Nakagawa and Mori, 2000), etc. Some methods rely purely on linguistic information, namely morpho-syntactic features of term candidates (Ananiadou, 1994). Hybrid approaches combining linguistic patterns and statistical measures (e.g. (Frantzi et al., 2000)) and machine-learning techniques (e.g. (Hatzivassiloglou et al., 2001)) have been also used.

However, few studies have been done for morphologically rich Slavic languages. For example, Vintar (2000) presented two methods for extraction of terminological collocations in order to assist the translation process in Slovene. The statistical approach was based on the mutual expectation and LocalMax measures, and involved collocation extraction from raw text. The extracted collocations were filtered with a stopword list, and only collocations containing single-word terms (devised previously by bilingual alignment) were accepted as relevant. In another approach, she used regular expression patterns to extract term collocations from a morpho-syntactically tagged corpus. However, these patterns are too general, and consequently not all extracted phrases were terminologically relevant.

In this paper we discuss automatic terminology recognition in Serbian, in particular, the extraction of multiword terms, which are very frequent[1] in certain domains (e.g. natural sciences, mathematics, etc.). Since Serbian is a highly inflective and morphologically and derivationally rich language, morpho-syntactic clues are indispensable in the ATR process. Our hybrid approach (called SRCE – Serbian C-value) combines morpho-syntactic features of term candidates and statistical analysis of their occurrences in text. In addition, since terms appear in texts in many different forms due to their morphological and derivational variations, the necessity of taking these variations into account becomes particularly apparent. Therefore, the SRCE method incorporates generic morpho-syntactic patterns, a term normalisation approach and a foreign word detection method.

The paper is organised as follows: in Section 2 we present an overview of the core term extraction method, called the C-value method. In Section 3 we discuss morpho-syntactic clues, the normalisation approach and the foreign word recognition that are used for singling out terms in Serbian. The experiments and evaluation are described in Section 4.

2 Automatic Term Recognition: the core C-value method

Our approach to ATR is based on the C-value method (Frantzi et al., 2000), which extracts multi-word terms. It is a general term recognition approach in the sense that it is not limited to specific classes of concepts. The approach is hybrid: the method combines linguistic knowledge (term formation patterns) and statistical analysis. Linguistic knowledge is used to single out term candidates, while their statistical features are used to measure the likelihood of term candidates being "real" terms. The method uses a POS tagged text as input, and outputs a list of extracted terms ranked according to their termhoods. *Termhood* is a numeric estimation of the degree to which a given linguistic unit (a multiword compound) is related to a domain-specific concept. However, the values are not normalised in the sense that a multiword, having a termhood value 10, is 10 times more likely to be a term than a term candidate with a termhood value 1.

In general, the C-value method enhances the commonly used baseline method that extracts most frequent term candidates (assuming that termhoods directly correspond to frequencies of occurrence) by making it sensitive to a particular type of terms — nested terms[2].

The method is implemented as a two-step procedure. In the first step, *term candidates* are extracted using a set of morpho-syntactic filters, which describe general term formation patterns in a given language. As a rule, terms form a proper subset of noun phrases (NPs). For example, a set of general filters for English may include the following patterns:[3]

Noun$^+$ Noun
(Adj | Noun)$^+$ Noun
(Adj | Noun)$^+$| ((Adj | Noun)* Prep?) (Adj | Noun)* Noun

Although these patterns are regular expressions, the filters are implemented as unification-like LR(1) rules (Mima et al., 1995) in order to facilitate processing of grammatical agreements (if any) within term candidates.

For each term candidate extracted by a filter, a set of nested term candidates is generated (see Table 1 for an example in English). The procedure for the generation of nested term candidates is implemented via transformation rules for each morpho-syntactic filter that is used to extract

[1] In English, more than 85% of domain-specific terms are multi-words (Nakagawa and Mori, 2000).

[2] For example, *nuclear receptor* is a nested term in *hormone nuclear receptor*. Similarly, *baza podataka* (Engl. *database*) is a nested term in *ažuriranje baze podataka* (Engl. *update of database*).

[3] *Noun*, *Adj* and *Prep* denote POS tags that correspond to nouns, adjectives and prepositions respectively. These filters were used for ATR from newswire corpora and in biomedicine (Frantzi et al., 2000; Nenadić et al., 2002).

term candidates. The main indicator that a nested term candidate might be a real term is that it also appears on its own in the corpus.

Term candidate:	Term
steroid hormone receptor factor	+
Nested term candidates:	
steroid hormone receptor	+
hormone receptor factor	-
steroid hormone	+
hormone receptor	+
receptor factor	-

Table 1: Nested term candidates

In the second step, the term candidates are assigned termhoods (referred to as *C-values*) according to a statistical measure. The measure amalgamates four numerical corpus-based characteristic of a candidate term, namely the frequency of occurrence, the frequency of occurring as nested within other candidate terms, the number of candidate terms inside which the given candidate term is nested, and the number of words contained in the candidate term. Formally,

$$C-value(a)=\begin{cases}\log_2|a|\cdot f(a), & a \text{ is not nested,}\\ \log_2|a|\left(f(a)-\frac{1}{|T_a|}\sum_{b\in T_a} f(b)\right), & \\ & \text{otherwise}\end{cases}$$

where a denotes a term candidate, $f(a)$ corresponds to its frequency, $|a|$ denotes the number of words in a, and T_a is a set of terms that contain term a as a nested term. Term candidates are ranked according to their C-values, and terms whose C-values are higher than a chosen threshold are presented as terms.

Evaluation of the C-value method for English has shown that using additional statistical information (frequency of "nestedness") improves the precision with slight loss on recall (Frantzi et al., 2000). Also, systematic term normalisation may further improve precision and recall of the method (Nenadić et al., 2002).

3 Morpho-syntactic clues for extraction of terms in Serbian

In order to adjust the core C-value method for Serbian, we have defined an appropriate set of morpho-syntactic filters and rules for inflectional normalisation of term candidates, and, additionally, a module for foreign word recognition.

3.1 Term formation patterns

As a rule, the vast majority of multiword terms in Serbian match the following general formation pattern:[4]

(1) $(Adj \mid ProAdj \mid Num \mid Noun)^{+}\ Noun$

which has been used for recognition of NPs in Serbian (Nenadić and Vitas, 1998a). Of course, not all NPs that follow this pattern are terms.[5] Moreover, when applied to an initially POS tagged text[6], this pattern may be too general even for description of NPs, as not all word sequences in a text that match this pattern are valid NPs. For example, in a sequence *koji se naziva relacioni model* (Engl. *which is called the relational model*), a word *naziva* can be initially tagged either as a noun *naziv* (Engl. *name*) or a verb *nazivati* (Engl. *call*), although, in this sentence, only the latter is correct. Thus, without further POS disambiguation, the string *naziva relacioni model* follows the pattern (1), although it is not a valid NP. This means that classical regular expressions are not sufficient for the representation of such constraints, and that we need more expressive means to model constraints related to the NP structure and agreements of multiword constituents on case, number and gender. We used the notion of generic patterns as an extension of regular expressions (Nenadić and Vitas, 1998b). For example, a generic pattern

(2) $Adj.x_1y_1z_1\ \ Noun.x_1y_1z_1\ \ Adj.x_2y_2g\ \ Noun.x_2y_2g$

models obligatory agreements that each NP from a specific class has to fulfil: both first and second pairs of adjectives and nouns must have the same values for certain morphological features (i.e. values for gender, number and case denoted by x_i,

[4] *ProAdj* and *Num* denote possessive adjectives and numbers respectively.

[5] For example, *ovaj način* (Engl. *this way), veliki deo* (Engl. *large part*), etc. This is a reason why we need additional processing to recognise semantically relevant NPs.

[6] Initially (or lexically) tagged POS text is a text in which every word occurrence is associated with *all* of its possible lexical and grammatical interpretations. The initial POS tagging is intrinsically ambiguous as each word is analysed separately, without considering neighbouring words (Nenadić and Vitas, 1998a). Thus, as a result of initial tagging, a lot of *lexical ambiguities* arise resulting in highly ambiguous word sequences. See Section 4 for further discussion.

y_i and z_i respectively), while these values may be different for each respective pair. The last adjective and noun are "frozen" in the genitive case (g), while the case (z_1) in the first pair is "free". By defining generic patterns one can model the agreements within various lexical structures in a highly inflective language such as Serbian (Nenadić and Vitas, 1998b). As a result, these agreements can be used to detect the boundaries of the structures in questions.

A set of generic patterns has been used to model the most frequent term formation patterns in Serbian. The set is mainly based on patterns used to model NPs in Serbian. Table 2 presents some of them. First four patterns describe NPs containing a nested NP whose lexical properties (such as case and/or number) are invariant in all inflected forms of the host NP. As a rule, the frozen part is in genitive. Depending on NP constituents, some agreements are obligatory within frozen part (see, for example, the third pattern – agreements between an adjective and the corresponding noun), or not (see the fourth pattern – no necessary agreement between the last two nouns in gender, number). The fifth pattern (Table 2) corresponds to NPs that do not have invariant parts.

Generic patterns		Examples
1	$N_1 \quad N_{gen}$	*baza podataka* *nejednakost trougla*
2	$A_1 \quad N_1 \quad N_{gen}$	*manipulativni aspekt modela* *granična vrednost niza*
3	$N_1 \quad A_{gen} \; N_{gen}$	*operacija prirodnog spajanja* *niz realnih brojeva*
4	$N_1 \quad N_{2;gen} \, N_{gen}$	*integritet baze podataka* *kriterijum konvergencije niza*
5	$A_1^+ \quad N_1$	*prošireni relacioni model* *kompletan metrički prostor*

Table 2: Frequent term formation patterns[7]

While these patterns are used to single out term candidates from an initially tagged text, agreements within NPs are used to generate possible nested structures. While the rules for nested structures are more "blurred" in English (since nouns are usually used as modifiers), "nestedness" in Serbian has to preserve the necessary structure and inner agreements, which are specific for the NP class in question. Therefore, generation of nested term candidates depends on the type of host term candidates (consider examples in Table 3). Nested structures that are not themselves NPs are not considered as term candidates.

Nested term candidates		NP	Term
2	*manipulativni aspekt modela* *manipulativni aspekt* *aspekt modela*	+ + +	+ - -
3	*operacija prirodnog spajanja* *operacija prirodnog* *prirodnog spajanja*	+ - +	+ - +
4	*integritet baze podataka* *integritet baze* *baze podataka*	+ + +	+ - +
5	*kompletan metrički prostor* *kompletan metrički* *metrički prostor*	+ - +	+ - +

Table 3: Nested term candidates (in Serbian)

3.2 Conflating morphological variants

If we aim at systematic recognition of terms, then handling term variation has to be treated as an essential part of terminology retrieval. Term variation ranges from simple orthographic (e.g. *oestrogen – estrogen*, *vitamin – vitamine*) and morphological variants (e.g. clone – clones) to more complex semantic variation (e.g. *eye surgery – ophthalmologic surgery*).

Several methods for term variation management have been developed. For example, the BLAST system (Krauthammer et al., 2000) used approximate text string matching techniques and dictionaries to recognise spelling variations in gene and protein names. FASTR (Jacquemin, 2001) handles morphological and syntactic variations by means of meta-rules used to describe term normalisation, while semantic variants are handled via WordNet.

The necessity of taking term variants into account as part of ATR process becomes particularly apparent in highly inflective languages. In Serbian, for example, the simplest morphological variations generally give rise to 14 possible variants of a single term (seven cases and two numbers (singular and plural) – see Table 4). If the

[7] In order to improve readability of filters, the generic patterns in this table are encoded using the following syntax: *A* and *N* stand for *Adj* and *Noun* respectively, while X_1 stands for $X.x_1y_1z_1$, X_{gen} stands for $X.xyg$ and $X_{2;gen}$ stands for $X.x_2y_2g$ (for $X \in \{A, N\}$). Also, invariant parts are underlined in the given examples.

core C-value method were to be applied without conflating morphological variants, then termhoods would be distributed across different morphological variants providing separate frequencies for individual variants instead of a single frequency calculated for a term candidate unifying all of its variants. In addition, the "nesting" factor of the C-value method would cause skewed results, since the case property of nested terms does not have normal distribution. Namely, as indicated previously (see Table 2), the majority of nested terms in Serbian are in genitive case, which means that the termhood for a term candidate in genitive case would differ significantly from its counterparts in other cases. Moreover, this deviation cannot be remedied later by summing up individual termhoods, since C-value is not an additive measure. Hence, in order for the C-value method to be applied correctly in a highly inflective language, term candidates must be (at least inflectionally) normalised *prior* to the calculation of termhoods.

Canonical form: *operacija prirodnog spajanja* (nom. sing. = ns)
Morphological variants: *operacija prirodnog spajanja* (ns;gp) *operacije prirodnog spajanja* (gs;np;ap;vp) *operaciji prirodnog spajanja* (ds;ls) *operaciju prirodnog spajanja* (as) *operacijo prirodnog spajanja* (vs) *operacijom prirodnog spajanja* (is) *operacijama prirodnog spajanja* (dp;ip;lp)
Normalised form: *operacija* (ns) *prirodno* (nsm) *spajanje* (ns)

Table 4: Variants and normalisation of term candidates – an example for term *operacija prirodnog spajanja* (Engl. *natural join operation*)

Our approach to morphological normalisation of term variants is based on the normalisation of individual term constituents. Namely, each word that is a part of a term candidate is mapped onto its lemma, and term candidates are treated as sequences of lemmas. At the end of the ATR process, terms are converted into their canonical form (singular, nominative case), which is not necessarily identical to the normalised form (the sequence of the corresponding singular words in singular, nominative case). The normalisation process is illustrated in Table 4.

At this point, the usage of generic patterns in order to check the agreements in case, number and gender during the phase of filtering of term candidates might seem unnecessary, since all these features are subsequently normalised. However, in order to enhance the precision of the SRCE method, it is important for term candidates to be correctly recognised prior to the statistical analysis. This means that the necessary agreements between NP constituents have to be checked. Once the term candidates are identified, they are normalised in order to make the most of the statistical part of the method.

3.3 Foreign word detection

Despite the efforts to rely mostly on Serbian vocabulary when building a terminology, many of the terms used in specific scientific domains borrow some of their building blocks from languages other than Serbian at various levels. For example, at morphological level, foreign suffixes, mostly originating from Latin and Greek, are often "preferred" to their Serbian counterparts in, for example, the biomedical domain, even when they are used to modify a root that is in fact Serbian (e.g. *amino-kiselina* (Engl. *amino acid*)). Similarly, at lexical level, words of foreign origin are used to form multi-word terms (e.g. *redundantan atribut* (Engl. *redundant attribute*)). This is particularly obvious in fairly recently expanded disciplines such as computer science, where, for many of the original terms used in English, it has not been simple to adapt new terms in Serbian. Consequently, many of the terms have been simply transcribed into Serbian or, even worse, they are still used in their original form. Not only do foreign words appear as "valid" parts of terms, but they have also proved to be good indicators of terms. It is, thus, necessary to develop procedures for their detection.

In our approach, the recognition of foreign words has been integrated into the ATR process for Serbian. The following morphological features are used to indicate occurrences of potential foreign words (Spasić, 1996):

- characters (e.g. *x*, *y*, *q*) that do not belong to Serbian graphemic system,
- successive vowel occurrences,
- exception to the palatalisation rule,

- exception to the assimilation rules,
- occurrence of atypical consonant bi/tri-grams
- occurrence of bi-grams or tri-grams typical for other languages (especially Latin and English), and
- foreign affixes.

The words satisfying some of the above criteria are not necessarily foreign words. The precision of these rules varies from one to another. For example, the first rule is the strongest indicator of the presence of foreign words, since the alphabetical system used is not Serbian. Other rules may be tuned to a certain extent in order to increase their precision.

Let us, for instance, consider the second rule. The successive usage of vowels is fairly frequent in Serbian, but the majority of such cases follow certain restrictions[8] under which they can occur. Moreover, these restrictions can be described by regular expressions. Any other occurrence of successive vowels can be used to indicate a potential foreign word.

Foreign word detection has been incorporated into the ATR process in two ways: during the selection of term candidates and for the calculation of termhoods. First, it is used before the initial POS tagging process in order to locate foreign words, which are tagged accordingly. Otherwise, foreign words would be typically considered as unknown. As explained earlier, it is very likely for foreign words in Serbian scientific and technical texts to be related to domain-specific concepts, and their mishandling would significantly decrease the recall of the ATR method. This information is used by the linguistic part of the SRCE-method, where we introduced a special category corresponding to foreign words.

In the second step, that is - once the term candidates have been selected - the information about foreign origin is used to increase the termhood of term candidates containing such words. This time, foreign word recognition is used to improve the precision of the ATR method.

4 Experiments and discussion

The preliminary ATR experiments were conducted using the SRCE system on a corpus containing samples from university textbooks in mathematics[9] and computer science[10] (altogether 120k words).

Texts were pre-processed, i.e. initially tagged, by a system of electronic dictionaries (e-dictionaries) containing simple nominal words for Serbian (Vitas, 1993). E-dictionaries contain exhaustive description of morpho-syntactic characteristics and are used for lexical recognition and initial lemmatisation of words that occur in a text. This process is realised by e-dictionary look-up, which results in an initially tagged text: each textual word is associated with its lemma(s) and corresponding morpho-syntactic categories (tags) retrieved from the e-dictionary. In general, e-dictionaries cannot resolve lexical ambiguities that result from the fact that there is no one-to-one correspondence between word forms and their morpho-syntactic features. There are different methods to resolve ambiguities (e.g. cache-dictionaries or local grammars), but in our experiments no disambiguation techniques were applied.

In order to extract a list of term candidates, the set of morpho-syntactic filters described in 3.1 was applied to the initially tagged corpus. We performed two sets of experiments.

In the first experiment, we did not use any stoplist to discard unwanted constituents of term candidates. For each term candidate, we generated a canonical form (nominative, singular), a morphologically normalised form (list of normalised words comprising the term candidate) and a list of nested term candidates (see Table 3 for examples). In the next step, C-values for term candidates were calculated using statistics based on occurrences of normalised forms, and all term candidates with C-values above an empirically chosen threshold were selected as terms.

Table 5 gives some examples of the recognised terms. In order to calculate the precision, we ex-

[8] For example, verbs in the paste tense, masculine gender always end with a pair of vowels (e.g. *ispitivao* (Engl. *examined*)). Further, some adjectives in masculine gender (e.g. *beo* (Engl. *white*)), as well as some nouns in masculine gender (e.g. *smisao* (Engl. *sense*)) also end with a pair of vowels. The usage of prefixes is another example where vowels may occur successively (e.g. *za+ustaviti* (Engl. *to stop*)).

[9] N. Lažetić, *Matematika II/1*, Naučna knjiga, Beograd, 1994

[10] G. Pavlović-Lažetić, *Osnove relacionih baza podataka*, Vesta - Matematički fakultet, Beograd, 1996. We would like to thank the authors of both textbooks for giving us permission to use their texts for experiments.

amined separately interval precisions in sub-corpora in mathematical analysis and computer science (see Table 6). Intervals are sets of recognised terms that are placed at certain positions within the list. For example, interval *1-50* contains top 50 terms, while the interval *over 150* contains all terms whose positions in the list are above 150. Terms have been inspected by the first two authors, who are Serbian native speakers and are specialists in both computer science and mathematics.

Term	C-value
metrički prostor	633.55
topološki prostor	175.13
otvoren skup	93.20
normiran prostor	88.00
Košijev niz	68.11
zatvoren skup	59.20
vektorski prostor	53.13
prirodan broj	44.41
nejednakost trougla	33.98
neprekidnost preslikavanja	28.02
Hausdorfov topološki prostor	19.43

Table 5: Top ranked terms in the domain of mathematical analysis

Interval	Mathematical analysis	Computer science
1 – 50	98%	90%
50 – 100	88%	70%
100 – 150	52%	58%
> 150	69%	68%

Table 6: Precision of the ATR method (without the usage of a stoplist)

In the first 50 terms for the domain of mathematical analysis, there was only one false term candidate (*specijalna klasa neprekidnih preslikavanja*), which contained an "unwanted" adjective *specijalna* (Engl. *special*). The reason for the significant drop in the precision in the second and third intervals is mainly the same: apart from few true negatives[11], the majority of false term candidates contained common "unwanted" constituents, which are sampled in Table 7. The results for the computer science sub-corpus were slightly worse since the mathematical language seems to be more consistent and restricted.

In the second experiment, we used a stoplist containing the words detected as frequent "wrong" constituents in the previous experiments. The results are summarised in Table 8.

prozvoljan	*opšti*	*pojam*
tražen	*dokazan*	*specifičnost*
specijalan	*globalan*	*svojstvo*
važan	*jedinstven*	*slučaj*
odgovarajući	*poznat*	*posledica*
definisan	*veliki*	*gledište*

Table 7: A sample of normalised stop-words

Interval	Mathematical analysis	Computer science
1 – 50	100%	94%
50 – 100	92%	92%
100 – 150	80%	74%
> 150	74%	70%

Table 8: Precision of the ATR method (with the usage of a stoplist)

The majority of remaining errors originate from the ambiguous POS tagging (more than 50%, problematic words being *naziv(a), igra, kod*, etc.). Since no further processing of text has been performed, another source of problems is the detection of boundaries of frozen parts in prepositional phrases (e.g. *na osnovu* (Engl. *based on*), *u slučaju* (Engl. *in the case of*)), which may be resolved by using a set of corresponding local grammars (Nenadić and Vitas, 1998b). In addition, for the computer science domain, some of the false terms were related to a specific application area (the text intensively used examples from a university information system, so candidates such as *zvanje nastavnika* (Engl. *lecturer position*), *godina studija* (Engl. *year of study*), etc. were wrongly suggested as computer science terms).

5 Conclusion

In this paper we have presented an approach to automatic extraction of terminology in a morphologically rich language, such as Serbian. Terms extracted automatically may be used as semantic indicators for a range of classic IR/IE tasks.

The approach is hybrid: it combines morpho-syntactic filters for extraction of term candidates, and statistical analysis that ranks term candidates according to their termhood.

[11] Such as: *toploška tačka gledišta, kompletnost prostora igra, kod preslikavnja*.

Extraction of term candidates is based on the recognition of proper NPs. In order to enhance both the precision and recall of the ATR method, it is inevitable to incorporate significant linguistic knowledge. Since describing NPs by means of regular expressions is not sufficient for modelling agreements between NP constituents, we have used generic morpho-syntactic patterns. Further, since not all NPs are terms that semantically characterise documents, we have used a statistical measure in order to estimate semantic significance of term candidates. Also, once the term candidates are correctly identified, they are normalised in order to make the most of the statistical part of the method. Term candidates suggested as terms by the statistical part of the SRCE method are finally mapped into the canonical form of the original term.

The preliminary experiments show that the precision is in line with the results for English, and that for the top ranked terms the precision is well above 90%. The analysis of errors shows that the majority of them appear due to lexical ambiguity of the input text. Certainly, if the corpora were lexically disambiguated, we would have better precision.

In order to improve the recall, additional morpho-syntactic filters need to be identified. In particular, we plan to study terms that contain prepositions, as this is a common formation pattern in many domains. Further, the broader handling of term variants (e.g. dialectic variants, acronyms, derivational variants) may also improve both precision and recall. Currently we deal only with inflectional variants by mapping them to a canonical form. Term variants unification and normalisation also provide a broader basis for further IR and IE tasks, as queries can be expanded by referring to a class of synonymous terms as opposed to a single term.

References

Ananiadou S. 1994. *Methodology for Automatic Term Recognition*. In Proceedings of COLING-94, Kyoto, Japan

Frantzi K.T., Ananiadou S. and Mima H. 2000. A*utomatic Recognition of Multi-word Terms: the C-value/NC-value Method*. Int. J. on Digital Libraries, 3/2, pp. 115-130.

Grobelnik M., Mladenić D. and Milić-Frayling N. 2000. *Text Mining as Integration of Several Related Research Areas*, KDD 2000 Workshop on Text Mining, Boston, USA

Hatzivassiloglou V., Duboue P. and Rzetsky A. 2001. *Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach*. Bioinformatics, 17/1, pp. S97-S106

Hearst M. 2000. *Text Mining Tools: Instruments for Scientific Discovery*, in IMA Text Mining Workshop, Institute for Mathematics and its Applications, Minneapolis, USA, 2000

Jacquemin C. 2001. *Spotting and discovering terms through NLP*. MIT Press, Cambridge MA, 378 p.

Krauthammer M., Rzhetsky A., Morozov P. and Friedman C. 2000. *Using BLAST for identifying gene and protein names in journal articles*. Gene, 259, pp. 245-252.

Mima H., Ando K. and Aoe J. 1995: *Incremental Generation of LR(1) Parse Tables*. In Proceedings of NLPRS'95, Pacific-Rim Symp., Seoul, Korea

Nakagawa H. and Mori T. 2000. *Nested Collocation and Compound Noun for Term Recognition*. Proc. of COMPUTERM 98, pp. 64—70

Nenadić G. and Vitas D. 1998a. *Formal Model of Noun Phrases in Serbo-Croatian*. BULAG 23, Universite Franche-Compte, Besançon, France.

Nenadić G. and Vitas D. 1998b. *Using Local Grammars for Agreement Modelling in Highly Inflective Languages*. In Proceedings of TSD 98. Masaryk University, Brno, pp. 91-96.

Nenadić G., Mima H., Spasić I., Ananiadou S. and Tsujii J. 2002. *Terminology-driven Literature Mining and Knowledge Acquisition in Biomedicine*. International Journal of Medical Informatics, 1-16.

Spasić I. 1996. *Automatic Foreign Words Recognition in a Serbian Scientific or Technical Text.* In Proceedings of Standardisation of Terminology, Belgrade, Yugoslavia, 1996

Vintar Š. 2000. *Extracting Terms and Terminological Collocations from the ELAN Slovene-English Parallel Corpus.* In Proceedings of the 5th EAMT Workshop, Ljubljana, Slovenia, 2000

Vitas D. 1993. *Mathematical Model of Serbo-Croatian Morphology (Nominal Inflection).* PhD thesis. Faculty of Mathematics, Belgrade.