Using English Baits to Catch Serbian Multi-Word Terminology

Cvetana Krstev, Branislava Šandrih, Ranka Stanković

Using English Baits to Catch Serbian Multi-Word Terminology

Cvetana Krstev, Branislava Šandrih, Ranka Stanković, Miljana Mladenović
Faculty of {Philology, Philology, Mining and Geology} University of Belgrade, College for Preschool Teachers
{Studentski trg 3, Studentski trg 3, Djušina 7} Belgrade, Bujanovac, Serbia,
cvetana@matf.bg.ac.rs, branislava.sandrih@fil.bg.ac.rs, ranka.stankovic@rgf.bg.ac.rs, ml.miljana@gmail.com

Abstract

In this paper we present the first results in bilingual terminology extraction. The hypothesis of our approach is that if for a source language domain terminology exists as well as a domain aligned corpus for a source and a target language, then it is possible to extract the terminology for a target language. Our approach relies on several resources and tools: aligned domain texts, domain terminology for a source language, a terminology extractor for a target language, and a tool for word and chunk alignment. In this first experiment a source language is English, a target language is Serbian, a domain is Library and Information Science for which a bilingual terminological dictionary exists. Our term extractor is based on e-dictionaries and shallow parsing, and for word alignment we use GIZA++. At the end of procedure we included a supervised binary classifier that decides whether an extracted term is a valid domain term. The classifier was evaluated in a 5-fold cross validation setting on a slightly unbalanced dataset, maintaining average F-score of 89%. After conducting the experiment our system extracted 846 different Serbian domain phrases, containing 515 Serbian phrases that were not present in the existing domain terminology.

Keywords: aligned texts, word alignment, terminology extraction, electronic dictionaries, morphological inflection

1. Motivation

Terminology is rapidly developing in many research and technological fields. It is very difficult to produce and maintain up-to-date terminology resources, especially for languages for which terminology in many fields is transferred and adapted from other languages. Such is the case for Serbian for which terminological resources in many domains, if existing, tend to be obsolete. Purely manual production of terminological resources is not the solution due to rapid changes both in research fields and corresponding terminology.

The work presented in this paper is motivated by our belief that Natural Language Processing (NLP) resources, methods and tools can help in the development of terminology in the Serbian language. Our work relies on the following presuppositions:

1. Serbian terminology is today transferred mainly from English because English terminology is better developed for many scientific and technological domains than Serbian (in the past from French and German). In (Ananiadou et al., 2012) lexical resources for English obtained grades 4.5–6 for all seven criteria, availability rated as excellent (the highest grade 6). To the contrary, the similar survey for Serbian (Vitas et al., 2012) showed that lexical resources are much less developed – they were rated 1–2.5.
2. Terminology consists mainly of Multi-Word Terms (MWT) (data presented in Subsection 4.2. corroborate this claim).[1]
3. A large portion of MWT terms in Serbian has a limited number of syntactic structures. Namely, 98% of all nominal MWEs in the Serbian general e-dictionary of MWEs has one of 13 different structures (having 2, 3 or 4 components) (Stankovic et al., 2016).

Under these presuppositions we are formulating the following hypothesis:[2]

> On the basis of the bi-lingual, aligned, domain-specific textual resources, the terminological list in the source language and the system for the extraction of terminology-specific nominal phrases (MWT) in the target language it is possible to compile the bilingual aligned terminological list.

2. Related Work

In recent years extraction of bilingual MWTs, and MWEs in general, from bilingual aligned corpora has been exploited by many researchers. Although most of them rely on automatic word alignment they differ both in resources and techniques used and in purpose for which they are compiled. In several cases the bilingual MWE lists are produced in order to improve statistical machine translation (Bouamor et al., 2012; Tsvetkov and Wintner, 2010) or to help developing certain lexical resources in the target language on the basis of the existence of such resource in the source language (e.g. used for the Slovenian WordNet (Vintar and Fišer, 2008)). In some cases, no lexical resources are used (Bouamor et al., 2012), while others rely on the existence of some bilingual lexicon (Tsvetkov and Wintner, 2010). MWEs are identified (in a source or a target language) in various ways: some authors use morphosyntactic patterns on lemmatized and POS-tagged texts

[1] Multiword expressions (MWE) are lexical units composed of more than one word, which are syntactically, semantically, pragmatically, and/or statistically idiosyncratic (Baldwin and Kim, 2010). MWTs are domain-specific MWEs.

[2] In this paper we will call 'source' language a well-resourced language (English), and 'target' language a less-resourced language (Serbian).

(Bouamor et al., 2012; Vintar and Fišer, 2008), while others perform full semantic parsing (Moirón and Tiedemann, 2006). For (Tsvetkov and Wintner, 2010) automatic word alignment is the main source of information for identifying MWEs.
In our case, we are relying on the existence of the lexical resource (terminology) in the source language. In order to align it with MWT in the target language we use neither full parsing (as not available for Serbian) nor we lemmatize and POS-tag text. Instead we use shallow parsing relying on extensive morphological e-dictionaries of Serbian (Cvetana Krstev, Duško Vitas, 2015) that not only helps to identify terminology precisely, but also enables production of correct MWT lemmas and consequently all its inflected forms (which is crucial for various applications, from searching to machine translation).

3. The Design of the System

The System consists of several components developed in C# and Python and interconnected to work in a pipeline. It relies on existing monolingual extraction of MWEs for Serbian implemented in LeXimir (Stankovic et al., 2016) and GIZA++ word alignment, while all other components are newly developed. The overall design of our system (Figure1) is as follows:

1. Input:
   - A sentence-aligned domain-specific corpus involving a source and a target language. We will denote an entry in this corpus with $S(text.align) \leftrightarrow T(text.align)$;
   - A list of terms from the same domain in a source language (both single-word terms (SWTs) and MWTs). We will denote an entry in this list with $S(term.list)$;
   - A list of MWTs extracted from the target part of the aligned corpus having some expected syntactic structure. We will denote an entry from this list with $T(term.extract)$.
2. Processing:
   - Aligning bilingual chunks (possible translation equivalents) from the aligned corpus. We will denote aligned chunks with $S(align.chunk) \leftrightarrow T(align.chunk)$;
   - Filtering the chunks to those in which a source part of a chunk matches a term from a list of domain terms in a source language: $S(align.chunk) \sim S(term.list)$, where symbol $\sim$ denotes the relation "match" (that is for our experiment defined in Subsection 4.5.);
   - Filtering once more previously filtered chunks to those in which a target part of a chunk matches a term from a list of extracted MWTs in a target language: $T(align.chunk) \sim T(term.extract)$;
3. The result: the list of filtered chunks that pass a certain threshold linked to matching source and target terms: $S(term.list) \leftrightarrow T(term.extract)$, where $(S(term.list) \sim S(align.chunk)) \wedge (T(term.extract) \sim T(align.chunk)) \wedge (S(align.chunk) \leftrightarrow T(align.chunk))$.

In order to test our approach and determine a threshold we have used the existing bi-lingual terminology resource in order to establish:

- How many aligned chunks after two-pass filtering have a target part that matches a target term in the bi-lingual resource: $(S(term.list) \leftrightarrow T(term.list)) \wedge (T(align.chunk) \sim T(term.list))$ (correspondence between terms in the bi-lingual terminology resource confirmed);
- How many aligned chunks after the first filtering have a target part that matches a term from the list of extracted terms in a target language and does not match a target term in the bi-lingual resource: $(T(align.chunk) \sim T(term.extract)) \wedge (T(align.chunk) \not\sim T(term.list))$ (a potentially new bi-lingual terminological terms);
- How many of aligned chunks obtained in the previous step contain a sound terminology in their target part (the new bi-lingual terminological pairs established).

On the basis of these results we developed a binary supervised classifier with aim to predict whether extracted terms belong to a domain terminology.

4. The Set-Up of the Experiment

In this section we will present resources that we used as an input for our experiment (subsections 4.1.– 4.4.), and the tools used in the processing steps (Subsection 4.5.). For this experiment we are using as an input:

- For the domain of Library and Information Science we have developed the English/Serbian textual resource containing 14,710 aligned sentences;
- Already mentioned Dictionary of Library and Information Science (English/Serbian pairs);
- The rule-based system for the extraction and lemmatization of potential terminological nominal phrases;
- Bilingual Serbian/English list of inflected word forms and MWE pairs derived from bilingual dictionaries and morphological (inflected) dictionaries for Serbian and English;

4.1. Aligned/parallel corpus

The English/Serbian textual resource was derived from the journal for Digital Humanities *Infotheca*[3] that is published biannually in Open Access. 12 issues with 84 papers were aligned at sentence level resulting in 14,710 alignment

[3] infoteka.bg.ac.rs/index.php/en

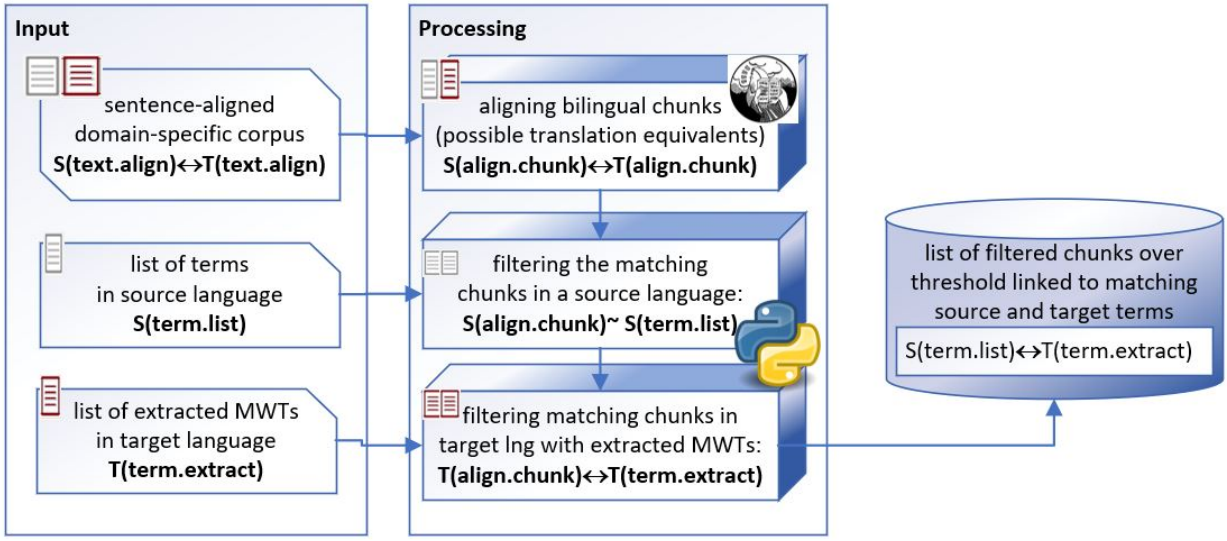

Figure 1: The overall design of the system for terminology extraction using monolingual and bilingual resources.

pairs. (Stanković et al., 2017); (Stanković et al., 2014).[4] The Serbian part has 301,818 simple word forms (41,153 different), while English part has 335,965 simple word forms (21,272 different). This means that in the Serbian part word forms repeat approximately 7 times, while in English they repeat 15 times. The major reason for this difference is the high inflection characteristic to Slavic languages producing many different forms for each lemma.

4.2. Dictionary of Library and Information Science

The development of the Dictionary of Librarianship: English-Serbian and Serbian-English (in this text referred to as 'Dictionary') (Ljiljana Kovačević, 2014) has started in 2001 at the National Library of Serbia, with the aim of presenting the librarianship terminology on different media (Kovačević et al., 2004).

This resource was first used on aligned texts in query expansion (Stanković et al., 2012); the Excel format of the dictionary was at that time transformed into a relational database. The version of the Dictionary that we used for our experiment has 12,592 different Serbian terms (9,376, 74% MWT), 11,857 different English terms (8,575, 72% MWT), and 17,872 distinct pairs.[5] Among distinct pairs in 10,574 cases both terms were MWT (60%), while in 1,923 cases a Serbian MWT had a single-word term equivalent in English (11%), and in 1,070 cases an English MWT had a single-word term equivalent in Serbian (6%). Both terms in a pair were SWTs in 4,305 cases (24%). Among Serbian SWTs, 1,378 are components of some MWTs, while the same occurs for 1,245 English SWTs. All important features of the original dictionary were preserved: the relations between translational equivalents, the synonymous relations within each specific language, and, important for Serbian, relations between Ekavian and Ijekavian pronunciation variants. For the research presented in this paper, we used only Ekavian variant (because texts were in this pronunciation) and only those English/Serbian translation pairs where at least one term in a pair is an MWT.

4.3. The rule-based system for the terminology extraction

The approach to terminology extraction for Serbian based on e-dictionaries and local grammars described in (Stankovic et al., 2016) was improved with additional syntactic patterns in order to cover as many terminology structures as possible (Ranka Stankovic, Cvetana Krstev, 2016). In this research, 20 syntactic structures grouped in 12 classes[6] for extraction purposes were used, which can extract the most frequent syntactic structures identified by an analysis of several Serbian terminological dictionaries and Serbian e-dictionary of MWEs.

This system was applied to the Serbian part of the aligned text (presented in 4.1.) and the results of its work are presented in Table 1.[7] For each class the syntactic structure it recognizes is output as well as the number of extracted forms and (word by word) lemmas.[8] The total number of

[4]Available for searching at *Biblisha* site `jerteh.rs/biblisha/Default.aspx`

[5]The version on the Web contains 40.000 entries (appr. 14.000 in Serbian, 12.400 in English and 14.000 in German) `http://rbi.nb.rs/en/home.html`

[6]One class can group MWTs with various syntactic structures (recognized by different graphs, or finite-state automata); all MWTs in one class have the same number and characteristics of components that inflect.

[7]A – adjective, N – noun, *g* – the genitive case, *i* – the instrumental case, Prep – preposition, *pc* – the case that agrees with the preceding preposition, x – a word separator or a MWU component that does not inflect; S_2 – two component MWU; S_3 – three component MWU; S_4 – four component MWU.

[8]One of the strongest features of this system is that it performs lemmatization of MWUs; however, for the purpose of this experiment, in order to ensure flexibility, we have done simple-word lemmatization. A proper lemmatization is left for the final phase. The difference between these two types of lemmatization can be illustrated with the following example: a simple-word

ID	Inflectional class	Syntactic pattern	Description	Forms	Lemmas
1	AxN	AxN	Both components inflect and agree in gender, number and case	18,249	13,948
2	2xN	2xN	1st component does not inflect	123	99
3	N2x	$NxN_{g(i)}$	1st component inflects; the second is always in gen/inst	7,724	6,528
4	NxN	NxN	Both components inflect and agree in case and number	3,775	3,522
5	N4x	$NxA_{g(i)}xN_{g(i)}$ $NxPrepxN_{pc}$	1st component inflects; the second and third are in gen/inst 1st component inflects; 3rd agrees in case with a preposition	9,647	8,909
6	AxN2x	AxN2x	1st component inflects and agrees with a structure 3.	2,662	2,480
7	AxAxN	AxAxN	All 3 components inflect and agree in gender, number and case	2,421	2,260
8	2xAxN	2xAxN	1st component does not inflect; 2nd and 3rd are structure 1	137	113
9	N6x	$Nx(S_3)_{g(i)}$ $NxPrep(S_2)_{pc}$	1st inflects; 2nd, 3rd and 4th are a structure 5–8 in gen/instr 1st component inflects; 3rd and 4th are a structure 1–4 in a case that agrees with a preposition	2,452	2,376
10	AxN4x	AxN4x	The 1st component inflects and agrees with a structure 5.	3,160	3,082
11	AxN6x	AxN6x	1st inflects and agrees with a structure 5–8 in gen/instr	1,263	1,252
12	N8x	$Nx(S_4)_{g(i)}$ $NxPrep(S_3)_{pc}$ $(S_2)xPrep(S_2)_{pc}$	1st inflects; remaining 4 are structure 8, 9 in gen/instr 1st inflects; last three are a structure 5–8 in a case that agrees with a preposition First two are structure 1-4; last two are a structure 1–4 in a case that agrees with a preposition	1,135	1,113

Table 1: Term candidates per inflectional classes

recognized forms is 52,748, counting same forms recognized by different finite-state automata (distinct 49,552). The total number of lemmatized forms is 45,682 (distinct 42,638).

4.4. Bilingual list of inflected word forms

We explored different ways of utilizing existing lexical resources to improve the quality of statistical machine alignment. In order to do that we have augmented the set of aligned sentences with inflected forms (English/Serbian). We have used two bilingual lexical resources. (a) Serbian Wordnet (SWN) (Cvetana Krstev, 2013) that is aligned to the Princeton WordNet (PWN)[9] and (b) a bilingual list containing general lexica with 10,551 English/Serbian entries. The production of the bilingual list of inflected forms was done in several steps:

1. First we compiled the parallel list from SWN and PWN containing 75,766 aligned English/Serbian literals. This list was merged with existing bilingual list yielding the new list of 86,317 entries.

2. To each Serbian noun, verb or adjective from the list compiled in the previous step we assigned its inflected forms obtained from the Serbian morphological dictionaries (Krstev, 2008). These inflected forms have various grammatical codes assigned to them that were used in the step 4.

3. We performed the similar procedure for English nouns, verbs and adjectives from the bilingual list. In order to obtain inflected forms with grammatical categories we used the English morphological dictionary from the Unitex distribution[10] and the MULTEX-East English lexicon.[11] Grammatical codes from these two sources were harmonized.

4. In the final step we aligned Serbian and English inflected word forms by using corresponding grammatical codes and harmonizing them as best as possible. In many cases one English word form was aligned with several inflected forms in Serbian. For example, the English noun *board* (in singular) is related to Serbian: *tabla*, *table*, *tabli*, *tablu*, *tablo*, *tablom*, while its plural form *boards* is related to *table*, *tabli*, *tablama*. MWEs from the bilingual list obtained in the first step were connected in the same way: for instance, *blotting paper* ↔ *upijajućeg papira*, *upijajućega papira*, *upijajućem papiru*, . . . and *blotting papers*↔ *upijajuće papire*, *upijajući papiri*, *upijajućih papira*, . . .

At the end of this procedure we obtained the bilingual list of inflected forms having 372,432 entries.

4.5. Alignment of Chunks

In order to acquire a list of aligned bilingual chunks several steps have to be performed. Our dataset included 14,710 aligned sentences, containing general lexica with 10,551 English/Serbian entries, parallel list from SWN and PWN containing 75,766 aligned English/Serbian literals and aligned Serbian and English inflected word forms having 372,432 entries (all described in previous subsections).

lemma of a multi-word form *bibliotečko-informacionom delatnošću* is *bibliotečki-informacioni delatnost*, while a correct lemma is *bibliotečko-informaciona delatnost* 'library and information activities'.

[9]Serbian WordNet can be browsed at `http://sm.jerteh.rs/`.

[10]Unitex/GramLab, a lexical-based corpus processing suite `http://unitexgramlab.org/`

[11]`https://www.clarin.si/repository/xmlui/handle/11356/1041`

First, our data had to be preprocessed by tokenization, truecasing and cleaning. In the next step the 3-gram translation model was built using KenLM (Heafield, 2011) followed by the training of the translation model. For the purpose of word-alignment, phrase extraction, phrase scoring and creating lexicalised reordering tables we used GIZA++,[12] (Och and Ney, 2003), together with *grow-diag-final* symmetrization heuristic (Koehn et al., 2003). Resulted phrase-table contained 1,672,362 potential translation equivalents. In order to discard as many as possible of those aligned pairs that are not exact translation of each other, two filtering steps were done. Each pair of aligned chunks from this list also contains information about inverse and direct phrase translation probability.[13] First we kept only those aligned chunks that have at least one of these probabilities greater than 0.85, simultaneously performing punctuation elimination and discarding those that consisted of punctuation and numbers only. This reduced the list to 982,598 aligned chunks.

For the second step we provided Bag-of-Words (BoW) representation for the English terms from the Dictionary (described at subsection 4.2.) and removed stop words from the list, so the list is mainly populated with domain words. Then we lemmatized each token from BoW using Natural Language Toolkit (nltk) Python library and its WordNet interface.[14] The same simple-word lemmatization was applied to the English parts of the aligned chunks. English parts of the aligned chunks that do not have at least one lemmatized content word present in the lemmatized BoW Dictionary representation were eliminated. Along with original and lemmatized form of chunks and their counts, we also kept information about translation order of words from the original phrase-table. This information helped us to make a backup of Serbian SWTs that translate as English SWTs. We eliminated these pairs, but also stored them in a separate file, so we can use them while obtaining final results (described throughly in Section 5., step 6). Final list contained 491,990 translation pair candidates.

We decided to enrich corpus with additional parallel lists (described in Subsection 4.4.) since we observed certain improvement in evaluations of translation quality. First we splitted corpus of aligned sentences into three disjoint parts: training (80%), development (10%) and test set (10%). BLEU score (Papineni et al., 2002) was obtained for the three different 3-gram language models. First model was trained only on the training set, tuned and tested, which obtained BLEU of 24.78. Second model was trained on the training set extended with the bilingual list containing the general lexica list and the parallel list from SWN and PWN. BLEU score for the test set increased to 24.93. Third model was extended with the list of inflected forms as in the real experiment setting and BLEU score increased to 26.21.

[12]Statistical Machine Translation toolkit can be found at `https://github.com/moses-smt/giza-pp`

[13]How phrase translation probabilities are determined can be read in more details at `http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases`

[14]More about Natural Language Toolkit for Python and its WordNet interface can be found at `http://www.nltk.org/howto/wordnet.html`

Before continuing to the next processing step, we defined "match" relation between chunks as follows: Let a chunk be represented as an unordered set of distinct words contained in it after stop words removal. Two chunks match if they have the same set representation of this kind.

5. Results and Evaluation

1. The number of distinct[15] aligned chunks after two-pass filtering (the English part matches some term in the Dictionary – $S(align.chunk) \sim S(term.dict)$ – and the Serbian part is a multi-word chunk, that is, it contains at least 2 content words) was 11,678.

2. In the next step the additional filtering was done – $(T(align.chunk) \sim T(term.list)) \wedge (T(term.list) \leftrightarrow S(term.dict)) \wedge (S(term.dict) \sim S(align.chunk))$, as a result 425 different MWTs from the Serbian Dictionary were matched with the Serbian part of the aligned chunk.

3. The aligned chunks from Step 1 were filtered with the additional condition $T(align.chunk) \sim T(term.extract)$. 2,266 chunks were obtained, 2,120 of them matching different extracted MWTs.

4. The aligned chunks from the step 2 were filtered with the additional condition $(T(align.chunk) \sim T(term.extract))$. 326 different Serbian MWTs were both matched with the Dictionary and extracted by the tool.

5. The aligned chunks from step 3 were filtered with the additional condition $(T(align.chunk) \not\sim T(term.list))$. 1,935 Serbian MWTs were extracted by our term extractor (they were not in the Dictionary; they, however, may be synonymous to some term already in the Dictionary due to the condition in step 1).

6. Among results obtained in the previous step there was a number of only partially correct pairs. Namely, some mostly simple-word English terms were aligned with Serbian MWTs that contain as a component the translation of English terms. An example of such situation is: the term LIBRARY translates as BIBLIOTEKA, but this Dictionary term is different from the Serbian chunk and Serbian extracted term (e.g. BIBLIOTEKA $\neq$ PARTNER BIBLIOTEKA 'participating library') and it was therefore kept in step 5. Most of these pairs were expelled in this step, reducing to 1,018 pairs at the end (see also Table 3.

There were 452 (44.4%) Serbian MWTs that have English Dictionary SWT as a translation, and 566 (55.6%) which have respective English Dictionary translation as an MWT. There were 575 extracted MWTs with frequency 1, 158 extracted MWTs with frequency 2 and 285 extracted MWTs

[15]Phrase table often contains several similar entries of the same phrase. For example, *at the digital library*, *for digital library*, *because digital library* and *of the digital library* would represent four different entries within phrase table. We observed these as one phrase, in the manner of the previously defined "match" relation.

with frequency greater or equal to three. For the sake of getting insight in the exact number of different new Serbian MWTs extracted during the last step, we asked professional from the librarianship domain to perform manual annotation of the extracted phrases. After manual validation, 515 extracted Serbian MWTs were evaluated as good translations of the paired English Dictionary phrases.

The examples illustrating this process are given in Table 3. MWTs extracted by the extractor that are equal to Serbian dictionary entries are determined in step 4 while those that are new are determined in step 6. New MWTs that represent only partial translations are deleted after step 5. Note that Serbian terms presented in this table are simple-word lemmatized, as explained in Subsection 4.3.. Correct MWE lemmatization is performed as a separate task.

At this moment we did not consider MWTs not recognized by our term extractor (condition $(T(align.chunk) \not\sim T(term.list))$ applied to the results of step 3 filtering), because they are mixed with false omissions due to enrichment of the aligned corpus with bilingual inflected dictionary (see Subsection 4.4.).

In order to make our system able to automatically decide whether an extracted term is a valid domain term, we trained a supervised binary classifier. All samples from the step 4 were considered good translations (331 samples) and this set was expanded with samples manually annotated as good translations during step 6 (515, total 846 positive samples). The remaining pairs from the step 6 were labeled as bad translations (negative class, 503 samples).

In the preprocessing step, we extracted 43 text features (also referred as "linguistic" features in (Ebert, 2017; Repar and Pollak, 2017)) from original (GIZA_SRP_ORIG) and lemmatized (GIZA_SRP_LEMM) form of Serbian chunk obtained from GIZA++, corresponding extracted Serbian term (SRP_EXTRACTED) and from the English part of the aligned chunk (GIZA_ENG_LEMM) and its Dictionary entry (ENG_DICT). These features are: 1) total number of words in Serbian and English chunks, extracted term and English Dictionary term (*_WC), 2) extracted term frequency (*_FREQ), 3) count of chunks in text (*_COUNT), 4) count of present diacritics in Serbian terms (*_DIACRITICS), 5) number of characters in English and Serbian terms (*_LEN), 6) ratio of diacritics count and length (*_DIACRITICS_LEN_RATIO) and 7) ratio of lengths of two different terms (*_LEN_RATIO). After eliminating high correlating or constant features, the final dataset contained 28 features.[16]

The performance of the classifier was evaluated in the 5-fold cross validation setting using the following metrics: accuracy (Acc), precision (P), recall (R) and F-score (F_1). After several different classifiers evaluation, Gradient Boost model (Friedman, 2001) implementation from the scikit-learn toolkit for Machine Learning for Python (Pedregosa et al., 2011) turned out to have the best performance on this dataset. Gradient boosting is an iterative technique that combines a set of weak learners and delivers improved prediction accuracy. The instances predicted correctly are given a lower weight and the ones misclassified are weighted higher, until best instance weights are found. The performance of our classifier per each fold k is displayed in Table 2.

k	1	2	3	4	5	Avg
Acc	0.844	0.848	0.870	0.852	0.888	0.861
R	0.906	0.899	0.912	0.865	0.896	0.896
P	0.856	0.864	0.897	0.896	0.908	0.884
F_1	0.881	0.881	0.905	0.880	0.902	0.890

Table 2: Gradient Boost classifier evaluation

Ten features with highest influence on the classification outcome are displayed in Figure 2.

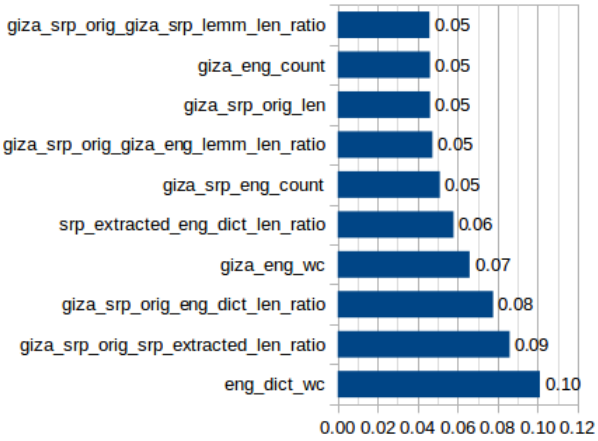

Figure 2: Influence of features on predictions

6. Conclusion

With this new experiment we wanted to achieve two goals: (a) to automatically evaluate extracted nominal phrases in the target language terminology by aligning it with the established terminology in the source language; (b) to build a classifier that would evaluate as positive automatically caught Serbian terminology using English baits. As an additional result we enriched the Dictionary of Library and Information Sciences with 515 synonyms in the Serbian part. Another by-product is the bilingual Serbian/English list of inflected word forms and MWE pairs derived from bilingual dictionaries and morphological dictionaries.

We will apply the same approach to other domains – mining, electro-distribution and management – since aligned domain corpora have already been prepared. At the same time the presented system will be improved with the user friendly interface for presentation of the results. Our intention is also to revise and further improve the relation "match" between aligned chunks and lexical resources, and possibly to introduce numeric values for the assessment rate. Needless to say, the enrichment of sentence-aligned domain-specific corpora is the long-term activity.

[16]The whole set of extracted features and the classifier itself are available on `https://github.com/Branislava/domain_terminology_extraction`.

Serbian Giza Chunk	Serbian Giza Lemmatized	Serbian Extracted	Serbian Dictionary	English Dictionary	Result sets	MWT
bibliotečkih udruženja sa	bibliotečki udruženje	bibliotečki udruženje	biblioteka društvo	association library	1, 3, 5, 6	yes
bibliotekarskim društvima	bibliotekarski društvo u	bibliotekarski društvo	biblioteka društvo	association library	1, 3, 5, 6	yes
bibliotečka asocijacija	nije bibliotečki asocijacija	bibliotečki asocijacija	biblioteka društvo	association library	1, 3, 5, 6	yes
udruženja i asocijacija poput biblioteke	udruženje i asocijacija poput biblioteka		biblioteka društvo	association library	1	
bibliotekarskog društva	bibliotekarski društvo	bibliotekarski društvo	bibliotekar	librarian	1, 3, 5	
bibliotečko informacionog sitema	bibliotečki informacioni sitema na	bibliotečki informacioni sistem	bibliotečki informacioni sistem	library information system	1, 2, 3, 4	
biblioteka medijateka	biblioteka medijateka		bibliotečki informacioni sistem	library information system	1	
drugim bibliotečko informacionim sistemima da	drugi bibliotečki informacioni sistem		bibliotečki informacioni sistem	library information system	1	
i informacioni sistemi u bibliotekama	informacioni sistem u biblioteka	informacioni sistem u biblioteka	bibliotečki informacioni sistem	library information system	1, 3, 5, 6	yes
koriste ovaj bibliotečko informacioni sistem	koji koriste ovaj bibliotečki informacioni sistem		bibliotečki informacioni sistem	library information system	1	
bibliotečke građe koju	bibliotečki građa koji	bibliotečki građa	bibliotečki građa	library material	1, 2, 3, 4	
biblioteka i bibliotečke građe	biblioteka i bibliotečki građa		bibliotečki građa	library material	1	
građe u bibliotekama	građa u biblioteka	građa u biblioteka	bibliotečki građa	library material	1, 3, 5, 6	yes
građu sve su biblioteke pod	građa sav su biblioteka		bibliotečki građa	library material	1	
bibliotečkog materijala	bibliotečki materijal i	bibliotečki materijal	bibliotečki građa	library material	1, 3, 5, 6	yes
građe iz drugih biblioteka a	građa iz drugi biblioteka		bibliotečki građa	library material	1	
bibliotečke standarde	bibliotečki standard	bibliotečki standard	bibliotečki građa	library material	1, 3, 5, 6	no
u pitanju bibliotečki materijal	pitanje bibliotečki materijal za		bibliotečki građa	library material	1	
biblioteka partnera	biblioteka partner i	biblioteka partner	biblioteka učesnica	participating library	1, 3, 5, 6	yes
biblioteke učesnice	biblioteka učesnica i	biblioteka učesnica	biblioteka učesnica	participating library	1, 2, 3, 4	yes
biblioteke učesnice i bibliotečki	biblioteka učesnica i bibliotečki		biblioteka učesnica	participating library	1	
su partneri biblioteke	partner biblioteka	biblioteka partner	biblioteka	library	1, 3, 5	
istraživanje iz oblasti bibliotekarstva i	istraživanje iz oblast bibliotekarstvo	istraživanje iz oblast bibliotekarstvo	naučni biblioteka	research library	1, 3, 5, 6	no
na sve naučne biblioteke	sav naučni biblioteka u		naučni biblioteka	research library	1	
of research libraries	of research libraries		naučni biblioteka	research library	1	
naučnoistraživačkih biblioteka	naučnoistraživački biblioteka u	naučnoistraživački biblioteka	naučni biblioteka	research library	1, 3, 5, 6	yes
naučne biblioteke ako su bile	naučni biblioteka ako su bile		naučni biblioteka	research library	1	
naučne biblioteke ako	naučni biblioteka u	naučni biblioteka	naučni biblioteka	research library	1, 2, 3, 4	
biblioteke ka istraživanju	biblioteka ka istraživanje		naučni biblioteka	research library	1	
iz istraživačkih biblioteka	istraživački biblioteka	istraživački biblioteka	naučni biblioteka	research library	1, 3, 5, 6	yes
knjiškomu moljcu	knjiški moljac	knjiški moljac	knjižni crv	bookworm	1, 3, 5, 6	yes
su razrešavanje višeznačnosti	razrešavanje višeznačnost i	razrešavanje višeznačnost	objašnjenje	disambiguation	1, 3, 5, 6	yes
za npr razrešavanje	npr razrešavanje		objašnjenje	disambiguation	1	
otklanjanje višeznačnosti su	otklanjanje višeznačnost	otklanjanje višeznačnost	objašnjenje	disambiguation	1, 3, 5, 6	yes
a razrešavanje	a razrešavanje		objašnjenje	disambiguation	1	
postoji povratna sprega	da postoji povratan sprega		povratan vez	feedback	1	
i povratne informacije	povratan informacija za	povratan informacija	povratan vez	feedback	1, 3, 5, 6	yes

Table 3: The examples illustrating the steps of terminology alignment

Acknowledgment

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178006.

7. Bibliographical References

Ananiadou, S., McNaught, J., and Thompson, P. (2012). *The English Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Ebert, S. (2017). *Artificial Neural Network Methods Applied to Sentiment Analysis*. Ph.D. thesis, Ludwig-Maximilians-Universität München.

Friedman, J. H. (2001). Greedy Function Approximation: a Gradient Boosting Machine. *Annals of statistics*, pages 1189–1232.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Kovačević, L., Injac, V., and Begenišić, D. (2004). *Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski*. Narodna biblioteka Srbije.

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Moirón, B. V. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word expressions in a multilingual context*, pages 33–40.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Repar, A. and Pollak, S. (2017). Good Examples for Terminology Databases in Translation Industry. In *eLex 2017: eLex 2017: The 5^{th} biennial conference on electronic lexicography, Netherlands, 19-21 September 2017*, pages 651–661.

Stanković, R., Krstev, C., Obradović, I., Trtovac, A., and Utvić, M. (2012). A tool for enhanced search of multilingual digital libraries of e-journals. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Stankovic, R., Krstev, C., Obradovic, I., Lazic, B., and Trtovac, A. (2016). Rule-based automatic multi-word term extraction and lemmatization. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Stanković, R., Krstev, C., sko Vitas, D., Vulović, N., and Kitanović, O., (2017). *Keyword-Based Search on Bilingual Digital Libraries*, pages 112–123. Springer International Publishing, Cham.

Tsvetkov, Y. and Wintner, S. (2010). Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1256–1264, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vintar, Š. and Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Vitas, D., Popović, L., Krstev, C., Obradović, I., zetić, G. P.-L., and Stanojević, M. (2012). *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

8. Language Resource References

Cvetana Krstev, Duško Vitas. (2015). *Serbian Morphological Dictionary - SMD*. University of Belgrade, HLT Group and Jerteh, Lexical resource, 2.0.

Cvetana Krstev. (2013). *Serbian WordNet*. University of Belgrade, HLT Group and Jerteh, Lexical database, 2.0.

Ljiljana Kovačević, Dr Dobrila Begenišić, Vesna Injac-Malbaša. (2014). *Dictionary of Library and Information Sciences*. National Library of Serbia: Scientific Research Department, Lexical database, 3.0.

Ranka Stankovic, Cvetana Krstev. (2016). *LeXimir*. University of Belgrade, HLT Group, Software Toolkit, 2.0.

Ranka Stanković and Cvetana Krstev and Nikola Vulović and Biljana Lazić. (2014). *Biblisha*. University of Belgrade, HLT Group, Bilingual digital library, 2.0.