Extraction of Bilingual Terminology using Graphs, Dictionaries and GIZA++

UDC 81'322.2
DOI 10.18485/infotheca.2019.19.2.6

ABSTRACT: In science, industry and many research fields, terminology is rapidly developing. Most often, a language that is "lingua franca" for most of these areas is English. As a consequence, for many fields, domain terms are conceived in English, and are later translated to other languages. In this paper, we present an approach for automatic bilingual terminology extraction for English-Serbian language pair that relies on an aligned bilingual domain corpus, a terminology extractor for a target language and a tool for chunk alignment. We examine the performance of the method on a Library and Information Science domain. The obtained results, as well as the application that implements the method, are available on-line.
KEYWORDS: terminology extraction, terminology validation, GIZA++, graphs, Unitex, text classification.

PAPER SUBMITTED: 30 September 2019
PAPER ACCEPTED: 20 December 2019

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs
University of Belgrade
Faculty of Philology

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology

Belgrade, Serbia

1 Introduction

In science, industry and many research fields, terminology is rapidly developing. Most often, a language that is "lingua franca" for most of these areas is English. As a consequence, for many fields, domain terms are conceived in English, and are later translated to other languages. It does not happen rarely that a certain term is translated either as a short explanation of its meaning, or the translation is specifically adapted as an utterance in the language in which it is translated to (i.e. as a word in a target language). An example that demonstrates both cases is an English word "a screenshot", from the computer science. In Serbian, this term is either translated as *snimak*

ekrana (namely, *a photo of a current state of the screen*) or as a "skrinšot" (i.e, the word is transcribed). It is not uncommon that even experts from a certain field have difficulties while translating texts that contain domain terminology. As in the example with a "debugger", the transcribed version is adopted for everyday use in Information Technologies domain.

It is a challenge to produce and maintain up-to-date terminology resources, especially for an under-resourced language, such as Serbian. Today, Serbian terminology is transferred mainly from English, since it is better developed for many scientific and technological domains. Purely manual production of terminological resources is not the solution due to rapid changes both in research fields and corresponding terminology.

Multi-Word Expressions (MWEs) are lexical units composed of more than one word, which are syntactically, semantically, pragmatically, and/or statistically idiosyncratic (Baldwin and Kim, 2010). MWEs represent a class of linguistic forms spanning conventional word boundaries that are both idiosyncratic and pervasive across different languages (Constant et al., 2017).

As Baldwin and Kim (2010), among others, have pointed out, the question of what constitutes a word is surprisingly complex, and one reason for this is the predominance of elements known as MWEs in everyday language. They consist of several words (in the conventionally understood sense) but behave as single words to some extent.

An illustration is given in (Constant et al., 2017), with a MWE *by and large*, that has roughly equivalent meaning and syntactic function to adverb *mostly*. Among the problematic characteristics of this expression are (1) syntactic anomaly of the part-of-speech (POS) sequence preposition + conjunction + adjective, (2) non-compositionality: semantics of the whole that is unrelated to the individual pieces, (3) non-substitutability of synonym words (e.g., by and big), and (4) ambiguity between MWE and non-MWE readings of a substring by and large (e.g., *by and large we agree* versus *he walked by and large tractors passed him*).

Due to all these difficulties, tackling MWEs represents a special challenge. This paper aims at MWEs since terminology consists mainly of Multi-Word Terms (MWTs). MWTs are domain-specific MWEs. Terms consisting of a single word are mainly referred to as Single-Word Terms (SMTs).

In this paper, we describe an approach for obtaining bilingual terminology pairs automatically, initially proposed in (Krstev et al., 2018) and demonstrated on English-Serbian language pair. In this first approach, we performed and discussed only one setting of the experiment. After evalua-

tion, we recognised a need to examine several settings of the experiment, which are conducted and discussed in the later text.

The proposed approach is based on the following hypothesis:

> On the basis of bilingual, aligned, domain-specific textual resources, a terminological list and/or a term extraction tool in a source language, and a system for the extraction of *terminology-specific Multi-Words Terms* in a target language, it is possible to compile a bilingual aligned terminological list.

This paper is organised as follows. An overview of previous work on this topic is given in Section 2. Lexical resources and tools that were used in the experiments in Subsection 3. The proposed approach is thoroughly explained in Section 4. Results and a discussion are given in Section 5. A Web application that implements the proposed technique is presented in Section 6. Finally, conclusions and directions for future work are given in Section 7.

2 Related Work

Over the past years, in order to compile bilingual lexica, researchers used various techniques for MWT extraction and alignment that differ in methodology, resources used, languages involved and purpose for which they were built.

Bilingual lexica were compiled for different language pairs: English/French (Bouamor et al., 2012; Hamon and Grabar, 2016; Hazem and Morin, 2016; Hakami and Bollegala, 2017; Semmar, 2018), English/Spanish (Oliver, 2017), English/Arabic (Lahbib et al., 2014; Naguib Sabtan, 2016; Hewavitharana and Vogel, 2016), English/Urdu (Hewavitharana and Vogel, 2016), English/Italian and English/German (Arcan et al., 2017), English/Slovene (Vintar and Fišer, 2008), English/Croatian, Latvian and Lithuanian (Pinnis et al., 2012), English/Chinese (Xu et al., 2015), English/Hebrew (Tsvetkov and Wintner, 2010), English/Ukrainian (Hamon and Grabar, 2016), English/Greek (Kontonatsios et al., 2014), English/Romanian (Pinnis et al., 2012; Kontonatsios et al., 2014), Bengali/Hindi/Tamil/Telugu (Irvine and Callison-Burch, 2016), Slovak/Bulgarian (Garabík and Dimitrova, 2015) and Italian-Arabic (Fawi and Delmonte, 2015).

In several cases, the bilingual lists of MWTs were compiled in order to improve statistical machine translation (SMT) of an existing machine translation system (Bouamor et al., 2012; Tsvetkov and Wintner, 2010; Naguib Sabtan, 2016; Irvine and Callison-Burch, 2016; Semmar, 2018; Hewavitharana and Vogel, 2016; Arcan et al., 2017; Oliver, 2017), for the development of an existing language resource in a target language on the basis of a corresponding resource in a source language (e.g. used for development of the Slovenian WordNet (Vintar and Fišer, 2008) based on English WordNet), or for the presentation of bilingual correspondences between two languages (e.g. correspondences between Slovak-Bulgarian parallel corpus (Garabík and Dimitrova, 2015)).

Some approaches request parallel sentence-aligned data (Arcan et al., 2017; Garabík and Dimitrova, 2015; Bouamor et al., 2012; Semmar, 2018), while others perform the extraction on comparable corpora (Xu et al., 2015; Hazem and Morin, 2016; Hewavitharana and Vogel, 2016; Pinnis et al., 2012). For the technique used in (Naguib Sabtan, 2016), groups of aligned sentences (verses) were used. In (Irvine and Callison-Burch, 2016) authors performed two experiments, the first one relying on the existence of a bilingual dictionary with no parallel texts and the second one requiring only the existence of a small amount of parallel data.

In order to compile a bilingual lexicon for a specific domain, we combined and compared several settings. Besides using only a parallel sentence-aligned corpus, we conducted an experiment where sentences from the corpus were extended with a bilingual list of inflected word forms from a general-purpose dictionary, similarly as in (Tsvetkov and Wintner, 2010).

We compared different configurations for the extraction of domain terminology on both, source and target, sides. For the source side, we compare two cases. In the first case, we use an existing bilingual domain dictionary, similarly as in (Vintar and Fišer, 2008; Hakami and Bollegala, 2017; Kontonatsios et al., 2014). In the second case, we obtain source terminology using an existing term extractor, similarly to some other authors (Pinnis et al., 2012; Hamon and Grabar, 2016; Arcan et al., 2017).

For the extraction of terminology on the target side, we apply morphological and statistical analysis. A similar approach was taken by other authors (Bouamor et al., 2012; Lahbib et al., 2014; Fawi and Delmonte, 2015; Hamon and Grabar, 2016; Naguib Sabtan, 2016; Semmar, 2018).

3 Lexical Resources and Tools

As previously mentioned in Section 1, the approach proposed in (Krstev et al., 2018) relies on several lexical resources and tools:

i A sentence-aligned domain-specific corpus involving a source and a target language, denoted as $S(text.align) \leftrightarrow T(text.align)$. In this paper we refer to this tool as LIS-CORPUS.
As a textual resource, twelve issues with a total of 84 papers were aligned at the sentence level resulting in 14,710 aligned segments (Stanković et al., 2017; Stanković et al., 2014).[1] The Serbian part has 301,818 simple word forms (41,153 different), while the English part has 335,965 simple word forms (21,272 different).

ii A list of terms in the source language, denoted as $S(term)$.
This list can be either an external resource from the same domain or extracted from the text.
As an external resource, we used the Dictionary of Librarianship: English-Serbian and Serbian-English. It was developed by a group of authors from the National Library of Serbia.[2] In this paper we refer to this tool as LIS-DICT.
We also tried to extract terms on the source side. For this purpose, we decided to use an open-source software tool, FlexiTerm (Spasić et al., 2013). It automatically recognises MWTs from a domain-specific corpus, based on their structure, frequency and collocations. In this paper we refer to this tool as ENG-TE.
Three other MWT extractors were considered for obtaining English MWTs: TextPro[3] (Pianta et al., 2008), TermSuite[4] (Cram and Daille, 2016) and TermEx2.8.[5] Evaluation performed on the list of terms extracted by all four extractors and evaluated as potential MWU terms showed that FlexiTerm outperformed the other three.

iii A list of terms in the target language, denoted as $T(term)$.

[1] Biblisha

[2] A more enhanced version of this dictionary, available on-line, contains 40.000 entries (approximately 14.000 in Serbian, 12.400 in English and 14.000 in German). We used the version obtained from the authors for research purposes.

[3] TextPro (former KX toolkit)

[4] TermSuite is the Open Source and UIMA-based application drawn out from the European project TTC Terminology Extraction

[5] TermEx

This list can be either an external resource from the same domain or obtained from the text.

The only system developed specifically for the extraction of MWTs from Serbian texts is a part of LeXimir (Stanković et al., 2016), a tool for management of lexical resources. LeXimir consists of two modules for the terminology extraction. The first module is a rule-based system relying on e-dictionaries and local grammars developed in Unitex,[6] that are implemented as finite-state transducers (FST). The second module implements various statistical measures used for ranking of term candidates. In this research the system was tuned to recognise 26 most frequent syntactic structures, which were previously identified by an analysis of several Serbian terminological dictionaries and the Serbian e-dictionary of MWUs (Krstev, 2008). In this paper we refer to this tool as Serb-TE. Some of these structures are `A_N_Prep_N` in *republički zavod za statistiku* 'republic office for statistics' or `A_N_(A_N)`$_{gen}$ in *statistički godišnjak republičkog zavoda* 'statistical yearbook of the republican institute' where `A` stands for an adjective, `N` for a noun and `PREP` for a preposition. Each of these components can be a single word or a MWU. Our system was used in a mode in which all possible MWTs in a word sequence are recognised, and not only the longest one. For instance, for the sequence *studija slučaja u primeni mašinskog učenja* 'case study in application of machine learning' the recognised terms would be: *studija slučaja* 'case study', *studija slučaja u primeni* 'case study in application', *mašinskog učenja* 'machine learning' and the longest match would be *studija slučaja u primeni mašinskog učenja*. The list of the most frequent classes is presented in (Krstev et al., 2018).

We have also prepared an additional resource, namely a set of aligned and inflected English-Serbian single and multi-unit word forms (denoted as bi-list). We used two bilingual lexical resources that we processed with LeXimir: (a) Serbian Wordnet (SWN),[7] which is aligned to the Princeton WordNet (Princeton WordNet, 2010), and (b) an English-Serbian list containing general lexica.

The production of bi-list was done in several steps:

1. First, a parallel list from SWN and PWN containing aligned English/Serbian Single and Multi-Word literals was compiled. This list was then merged with the bilingual list yielding a new list.

[6] Unitex/GramLab, a lexical-based corpus processing suite

[7] Serbian WordNet

2. To each Serbian noun, verb or adjective from the merged list we assigned its inflected forms obtained from the Serbian morphological e-dictionaries (Krstev, 2008). These inflected forms have various grammatical codes assigned to them, which were used in the final step.
   As mentioned earlier, 26 most frequent syntactic structures, grouped in 14 graphs labelled: AXN, 2XN, N2X, N4X, AXN2X, NXN, AXAXN, N6X, AXN4X, 2XAXN, AXN6X, 12N8X, 2XAXN2X and 2XAXN4X were used for terminology extraction. In this notation, A stand for and adjective, N for noun, X for a component that do not inflect. The separators like space and hyphen are also labeled as X and nX is short notation for repetition of X n times.
   As an example, we present a FST for extraction of type N4X, meaning that the first component is noun that inflects, followed by two words that do not inflect. The N4X graph is shown in Figure 1, which shows two paths that recognize two syntactic structures:
   - N_Prep_Np, noun (1^{st} component that inflects) followed by prepositional phrase (3^{rd} component agrees in case with a preposition), as in examples: *lista sa podacima* (list with data), *mašina za pranje* (washing machine), *ugovor o radu* (work contract);
   - N_Ngi_Ngi NxAg(i)xNg(i), 1^{st} component inflects; the second and the third component (noun or adjective) are in genitive case (such as *izrada geološke karte* (creation of a geological map)) or instrumental case (such as *etiketiranje vrstom reči* (Part-of-Speech tagging))

   The graph output consists of 4 values for each recognised MWU, separated by ";": graph label (grf04a or grf04b), followed by a label that indicates grammatical number (sin or plu); followed by recognised form ($n.INFLECTED$ p np or $n.INFLECTED$ $ng1$ $ng2$) and lemmatised inflective component followed by constant components ($n.LEMMA$ p np or ($n.LEMMA$ $ng1$ $ng2$). An example would be: "grf04b;plu;ciljeva pronalaženja informacija;cilj pronalaženja informacija;" (goal of finding information).
   A Software solution for multi-word units extraction displayed in Figure 2 offers possibilities for general NLP processing on selected corpus (applying lexical resources, generating bag of words and extraction of unknown words), extraction of selected syntactic patterns applying specified options and further processing (lemmatisation, calculation of statistical measures, support for manual evaluation and final evaluation report). For automatic extraction and lemmatisation, the system calls Unitex command-line functions in the background to apply appropriate graphs.

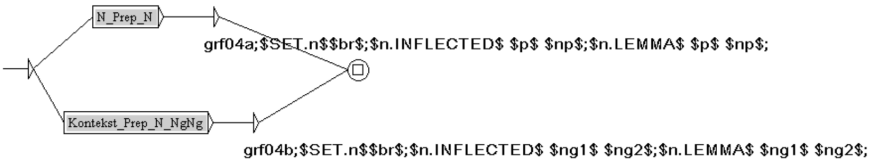

Figure 1. A FST for extraction of type N4X

3. A similar procedure was performed 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 and the MULTEX-East English lexicon.[8]
4. In the final step Serbian and English inflected word forms were aligned taking into account the corresponding grammatical codes, which were previously harmonised to the best possible extent.
   For example, the grammatical category codes in the Serbian dictionary are a/b/c, for the positive/comparative/superlative forms. The Unitex English dictionary does not have a code for the positive, while the codes for the comparative and superlative are C and S, respectively. The second English dictionary followed the MULTEXT-EAST specification, using p/c/s as codes. Thus the Serbian codes a/b/c were mapped to English codes ϵ/C/S and p/c/s, respectively.

4 Terminology Extraction

In our experiments the source language is English, and the target language is Serbian. For input and processing we used resources and tools described in Section 3. As the aligned corpus (Input i) we used LIS-CORPUS alone, or augmented with bilingual pairs from the BI-LIST. For the extraction of English terms (Input ii) we used the English side of the dictionary LIS-DICT in one series of experiments, and term extractor ENG-TE in the other, while the extraction of Serbian terms (Input iii) was done by SERB-TE.

With the notation introduced in Section 3, the extraction procedure consists of the following steps:

[8] MULTEX-East English lexicon

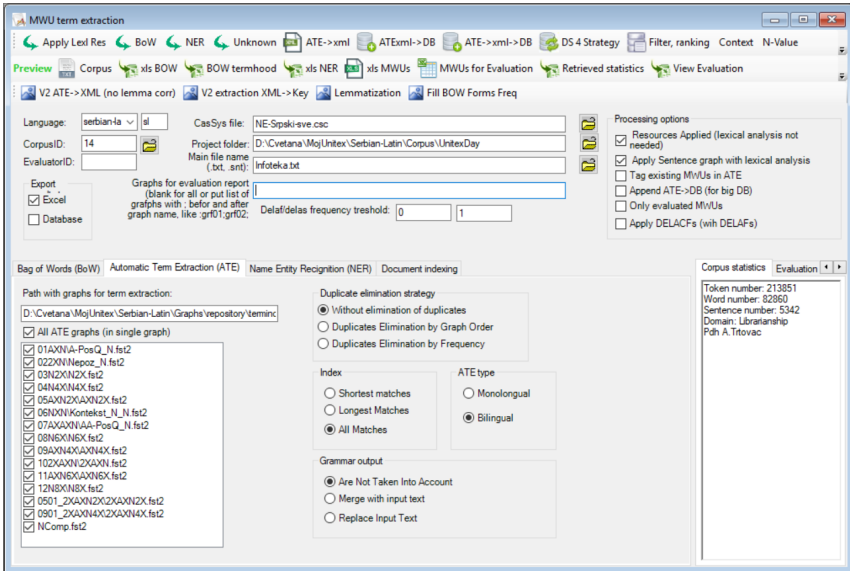

Figure 2. Software solution for MWT extraction

i Aligning bilingual chunks (possible translation equivalents) from the aligned corpus. We will denote aligned chunks by $S(align.chunk) \leftrightarrow T(align.chunk)$.

The alignment of chunks began with pre-processing using MOSES (Koehn et al., 2007) to perform tokenisation, truecasing and cleaning. In the next step a 3-gram translation model was built using KenLM (Heafield, 2011), followed by the training of this translation model. For the purpose of word-alignment, phrase extraction, phrase scoring and creation of lexicalised reordering tables, GIZA++[9] (Och and Ney, 2000) was used, together with the *grow-diag-final* symmetrisation heuristic (Koehn et al., 2003).

Each pair of aligned chunks from this list also contained information about inverse and direct phrase translation probability.[10] We have initially discarded all aligned chunks that did not have at least one of these probabilities greater than 0.85, simultaneously eliminating punctuation

[9] Statistical Machine Translation toolkit

[10] The way phrase translation probabilities are determined

marks. Chunks that consisted of punctuation marks and digits only were also discarded.

Afterwards, we provided a Bag-of-Words (BoW) representation for English terms from the LIS-DICT, i.e. from ENG-TE, and removed stop words from it, producing a list mainly populated with content words. Then we lemmatised each token from the BoW. Aligned chunks in which the English part did not have at least one lemmatised content word from the BoW list were eliminated.

ii Keeping only chunks (from the previous step) in which the source part of the chunk matches a term in the list of domain terms in the source language remain: $S(align.chunk) \sim S(term) = \{(s_1, s_2) : s_1 \sim s_2\}$, where the symbol $\sim$ denotes the relation "match" (explained later).

iii Keeping only chunks (from the previous step) in which the target part of the chunk matches a term in the list of extracted MWTs in the target language remain: $T(align.chunk) \sim T(term) = \{(t_1, t_2) : t_1 \sim t_2\}$.

The relation "match" ($\sim$) is defined as follows: if a chunk is represented by an unordered set of distinct words obtained from the chunk after removal of stop words, the two chunks match if they are represented by the same set. For example, if there is one "dictionary words" chunk and another "words from dictionary" chunk, their corresponding set representations are {dictionary, words} and {words, dictionary}, respectively ('from' should be discarded as functional word). Since these two sets are equal, these two chunks match.

Let two candidate pair chunks be "reči iz rečnika" (translated as 'words from dictionary') and "reč o rečniku" (translated as 'dictionary words'). Considering the specific application, these two chunks should match. If observed as unordered set of distinct content words, these chunks can be written as {reči, rečnika} and {reč, rečniku} ("iz" and "o" are prepositions, meaning *from* and *about*, and should be discarded as a functional word). Conceived like this, these two sets are different. For the best possible matching, chunks have to be normalised. This especially applies for highly inflectional languages, such as Serbian. In this specific case, Simple-Word lemmatisation within MWTs is needed. This means that each word from a MWT has to be replaced by a corresponding lemma from the available morphological e-dictionaries for Unitex (Krstev, 2008). For example, a word "reči" is a noun, has feminine gender, is in plural and is in nominative case. A lemma for any noun is singular, and is nominative case, namely "reč" for this case. The words "rečnika" and "rečniku" are also both nouns, but in genitive and dative case, respectively. After single-word lemmatisation, both of these words

are replaced with their lemma "rečnik". After this lemmatisation, for both chunks, set representations are {reč, rečnik} and {reč, rečnik}, and they, therefore, match.

A list of the resulting matched source and target terms $S(term) \leftrightarrow T(term)$, obtained from the aligned chunks, was retrieved as:

$$S(term) \leftrightarrow T(term) = \{(s,t) :$$
$$s \in S(term) \sim S(align.chunk) \wedge$$
$$t \in T(term) \sim T(align.chunk) \wedge$$
$$(s,t) \in (S(align.chunk) \leftrightarrow T(align.chunk)\}$$

5 Results and Discussion

The input preparation steps as well as processing consist of several components developed in C# and Python that are interconnected to work in a pipeline. The pipeline relies on existing tools for the extraction of English MWTs (ENG-TE) and Serbian MWEs (SERB-TE) implemented in LEXIMIR (Stanković et al., 2016) and on GIZA++ for word alignment, while all other components are newly developed.

In our experiments we combined each of the three following parameters, all related to the preparation of the input, where each parameter comes in two options, thus obtaining 8 different experimental settings:

1. The input domain aligned corpus (Input i) consists of:
   (a) the aligned corpus LIS-CORPUS;
   (b) the aligned corpus LIS-CORPUS extended with the bilingual aligned pairs BI-LIST (LIS-CORPUS+);
2. The list of domain terms for the source language (Input ii) is
   (a) the source language part of LIS-DICT including SWTs;
   (b) the output of the extractor ENG-TE applied to the source language part of the aligned input corpus;
3. The extraction of the set of MWTs in the target language by SERB-TE (Input iii) was done:
   (a) on the target language part of the aligned chunks (CHUNK);
   (b) on the target language part of the aligned input sentences (TEXT).

The summary of results obtained by our system for 8 experiment settings is given in Table 1. We refer to the experiments using the labels introduced above.

The numbers in the columns represent the following results:

Input and GIZA++ output results

A Number of entry pairs in LIS-DICT, i.e. English terms extracted by ENG-TE;

B Number of lines obtained from GIZA++ phrase table, after preprocessing steps;

C Number of distinct, lemmatised Serbian MWTs extracted from the target language part of the aligned chunks (for CHUNK) or from the target language part of the aligned input corpus (for TEXT).

Table 1. Numerical data that describes the results of the term extraction system

Experiment			A	B	C	I	II	III	IV
LIS-DICT	LIS-CORP	CHUNK	17,889	240,253	26,719	6,646	1,141	647	173
		TEXT			49,632		1,531	770	240
	LIS-CORP+	CHUNK		496,787	45,813	11,740	2,508	1,105	301
		TEXT			50,644		2,500	1,075	362
ENG-TE	LIS-CORP	CHUNK	3,169	215,317	35,226	5,063	2,233	x	x
		TEXT			49,632		2,233	x	x
	LIS-CORP+	CHUNK		446,979	44,885	8,164	3,333	x	x
		TEXT			50,644		3,310	x	x

Additional filtering of results obtained by GIZA++[11]

I Number of the aligned chunks after initial filtering using English terms (Processing ii): $(S(align.chunk) \sim S(term))$, where the list of English terms depends on the choice of parameter 1 (the English part of LIS-DICT or obtained from the corpus by using ENG-TE for extraction).

II Number of aligned chunks after subsequent filtering using Serbian terms (Processing iii) : $(S(term) \sim S(align.chunk)) \wedge (T(term) \sim T(align.chunk)) \wedge (S(align.chunk) \leftrightarrow T(align.chunk))$.

[11] To keep it simple, in the following notation, we refer to sets as to single representative terms, e.g. when we write $S(align.chunk)$, we refer to one term from that list.

III Number of new term pairs after filtering, namely those that do not already exist in LIS-DICT — these term pairs were obtained by selecting filtered chunks in which the Serbian part of the chunk does not match a term in the Serbian part of LIS-DICT ($(T(align.chunk) \not\sim T(term.list))$) (applicable only when LIS-DICT is used in the experiment);

IV Number of term pairs after filtering that already exist in LIS-DICT – these term pairs were obtained by selecting filtered chunks in which the Serbian part of the chunk matches a term in the Serbian part of ($T(align.chunk) \sim T(term.list)$) (also applicable only for (LIS-DICT) experiments).

In order to assess the efficiency of our approach, we have first evaluated all extracted pairs manually. Evaluation results showed that a number of new term pairs were retrieved. When LIS-DICT was used as a source of English terminology, 364 English terms from the dictionary were linked to new Serbian translations yielding 428 new term pairs. Among all term pairs retrieved using ENG-TE for extraction, 538 were supported by LIS-DICT, while among all term pairs retrieved using LIS-DICT for extraction, 168 were also retrieved with ENG-TE. A detailed evaluation procedure and results were described in (Šandrih et al., 2019).

6 BiLTe Web Application

In this Section, a Web application[12] that implements the proposed technique for terminology extraction is presented. The tool is freely available for online use.

The Web application consists of three modules: 1) input, 2) alignment and post-processing and 3) results module. Each module is briefly described and shown in the following Subsections.

Input Module

First, a user has to upload two sentence-aligned text files. Files must have the same names. File extensions should differ and indicate language (e.g. medicine.en and medicine.sr). These files are later fed into GIZA++.

Afterwards, a user has to upload a list of English terms. The first line should contain a header, and each line should contain one term.

[12] BILingual Terminology Extraction

Finally, a user has to upload a list of terms in Serbian (not necessarily MWUs). The first line is a header, each line contains a term and its frequency (for filtering later), separated with | ("pipe" character).

The interface of this module is displayed in Figure 3.

Figure 3. Input module of the BiLTe Web application

Alignment and Post-Processing Module

Aligning with GIZA++ yields a so called "phrase-table".

The alignment works in the following way. GIZA++ reads the two input texts in parallel. Whenever two bilingual chunks appear together, their co-occurrence is written into text file (dubbed *f_phrases*). Afterwards, *f_phrases* is sorted in two ways (by the target term and by the source term), and that's how two tables are obtained.

As earlier mentioned in Section 4, in step (ii), we discard candidates with direct or inverse probabilities lower than the threshold. After this, two post-processing steps follow. The first step is filtering by discarding terms that are out of the domain. This step is followed by a lemmatisation of English chunks with WordNet (Princeton WordNet, 2010) and Serbian chunks with e-dictionaries for Serbian (Krstev, 2008) (Procedure, i).

The interface of this module is displayed in Figure 4.

Figure 4. Pre-processing and Alignment module of the BiLTe Web application

Results Module

The basic steps of this module are: 1) keeping only candidates present in the English list (Procedure, ii), 2) performing intersection with Serbian extracted MWUs (Procedure, iii) and 3) additional filtering (optional) of bad candidates from the previous step.

The interface of this module is displayed in Figure 5.

Figure 5. The module for obtaining results of the BiLTe Web application

7 Conclusion

We conclude that the best results, in terms of quantity and quality of the obtained pairs, were achieved when input sentences were enhanced with additional bilingual pairs, and when extraction of Serbian terms was performed on the Serbian part of the aligned corpus, instead of aligned chunks. We will continue to experiment with these settings. Moreover, we intend to enrich BI-LIST with newly produced pairs. Our experiments also show that both methods of extraction produce some different pairs of equivalent terms. In our future work we will use not only both methods, when a dictionary for a source language becomes available, but also terms obtained from several different extractors. Another indented work is the integration of lemmatisation procedure into the bilingual extraction, already developed and implemented in monolingual MWU extraction, as described in (Stanković et al., 2016).

We intend to apply the same approach to other domains — mining, electric power system and management — for which aligned domain corpora have already been prepared. Of course, the enrichment of sentence-aligned domain-specific corpora, bilingual word lists and monolingual dictionaries of MWTs are long-term activities.

Acknowledgment

This research was partly supported by the Ministry of Education, Science and Technological Development through projects ON-178006 and III47003.

References

Arcan, Mihael, Marco Turchi, Sara Tonelli and Paul Buitelaar. "Leveraging Bilingual Terminology to Improve Machine Translation in a Computer Aided Translation Environment". *Natural Language Engineering* Vol. 23, no. 5 (2017): 763–788

Baldwin, Timothy and Su Nam Kim. "Multiword Expressions". *Handbook of Natural Language Processing* Vol. 2 (2010): 267–292

Bouamor, Dhouha, Nasredine Semmar and Pierre Zweigenbaum. "Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation". In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard et

al.. Istanbul, Turkey: European Language Resources Association (ELRA), 2012

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch et al. "Multiword Expression Processing: A Survey". *Computational Linguistics* Vol. 43, no. 4 (2017): 837–892

Cram, D. and B. Daille. "Terminology Extraction with Term Variant Detection". In *Proceedings of ACL-2016 System Demonstrations*, 13–18, 2016.

Fawi, F. and R. Delmonte. "Italian-Arabic Domain Terminology Extraction From Parallel Corpora". In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, Vol. 130, Accademia University Press, 2015

Garabík, Radovan and Ludmila Dimitrova. "Extraction and Presentation of Bilingual Correspondences from Slovak-Bulgarian Parallel Corpus". *Cognitive Studies | Études cognitives* no. 15 (2015): 327–334

Hakami, H. and D. Bollegala. "A Classification Approach for Detecting Cross-lingual Biomedical Term Translations". *Natural Language Engineering* Vol. 23, no. 1 (2017): 31–51

Hamon, T. and N. Grabar. "Adaptation of Cross-lingual Transfer Methods for the Building of Medical Terminology in Ukrainian". In *Proceedings of the* 17^{th} *International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2016)*, LNCS. Springer, 2016

Hazem, Amir and Emmanuel Morin. "Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora". In *Proceedings of COLING 2016, the* 26^{th} *International Conference on Computational Linguistics: Technical Papers*, 3401–3411, 2016.

Heafield, Kenneth. "KenLM: Faster and Smaller Language Model Queries". In *Proceedings of the* 6^{th} *Workshop on Statistical Machine Translation*, 187–197. Association for Computational Linguistics, 2011.

Hewavitharana, Sanjika and Stephan Vogel. "Extracting Parallel Phrases from Comparable Data for Machine Translation". *Natural Language Engineering* Vol. 22, no. 4 (2016): 549–573

Irvine, Ann and Chris Callison-Burch. "End-to-end Statistical Machine Translation with Zero or Small Parallel Texts". *Natural Language Engineering* Vol. 22, no. 4 (2016): 517–548

Koehn, Philipp, Franz Josef Och and Daniel Marcu. "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*, Vol. 1, 48–54. Association for Computational Linguistics, 2003.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico et al.. "Moses: Open Source Toolkit for Statistical Machine Translation". In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics, 2007

Kontonatsios, G., M. Claudiu, Korkontzelos I., Thompson P and S. Ananiadou. "A Hybrid Approach to Compiling Bilingual Dictionaries of Medical Terms from Parallel Corpora". *Statistical Language and Speech Processing* Vol. 8791 (2014): 57–69

Krstev, Cvetana. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade, 2008. https://hal.archives-ouvertes.fr/hal-01011806

Krstev, Cvetana, Branislava Šandrih, Ranka Stanković and Miljana Mladenović. "Using English Baits to Catch Serbian Multi-Word Terminology". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, chair), Nicoletta Calzolari (Conference, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al., 7–12. Paris, France: European Language Resources Association (ELRA), 2018. http://www.lrec-conf.org/proceedings/lrec2018/pdf/384.pdf

Lahbib, W., I. Bounhas and B. Elayeb. "Arabic-English Domain Terminology Extraction from Aligned Corpora". In *On the Move to Meaningful Internet Systems (OTM 2014) Conferences, Confederated International Conferences : CoopIS, and ODBASE 2014, Amantea, Italy, October 27-31, 2014, Proceedings*, Robert Meersman, Tharam Dillon Michele Missikoff Lin Liu Oscar Pastor Alfredo Cuzzocrea & Sellis Timos, Hervé Panetto, 745–759. Springer Berlin Heidelberg, 2014,

Naguib Sabtan, Yasser Muhammad, "Bilingual Lexicon Extraction from Arabic-English Parallel Corpora with a View to Machine Translation". *Arab World English Journal* Vol. 7, no. 5 (2016): 317–336. http://search.ebscohost.com.proxy.kobson.nb.rs:2048/login.aspx?direct=true&db=edb&AN=115896070&site=eds-live

Och, Franz Josef and Hermann Ney. "Improved Statistical Alignment Models". In 38th *Annual Meeting on Association for Computational Linguistics*, 440–447. Stroudsburg, PA: Association for Computational Linguistics, 2000.

Oliver, Antoni. "A System for Terminology Extraction and Translation Equivalent Detection in Real Time: Efficient use of Statistical Machine

Translation Phrase Tables". *Machine Translation* Vol. 31, no. 3 (2017): 147–161

Pianta, E., C. Girardi and R. Zanoli. "The TextPro Tool Suite". In *Proceedings of 6th edition of the Language Resources and Evaluation Conference*, 2008

Pinnis, Marcis, Nikola Ljubešic, Dan Stefanescu, Inguna Skadina, Marko Tadic et al.. "Term Extraction, Tagging, and Mapping Tools for Under-resourced Languages". In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, 20–21. 2012

Princeton WordNet, 2010

Semmar, Nasredine. "A Hybrid Approach for Automatic Extraction of Bilingual Multiword Expressions from Parallel Corpora". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, chair), Nicoletta Calzolari (Conference, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al.. Paris, France: European Language Resources Association (ELRA), 2018

Spasić, Irena, Mark Greenwood, Alun Preece, Nick Francis and Glyn Elwyn. "FlexiTerm: a Flexible Term Recognition Method". *Journal of Biomedical Semantics* Vol. 4, no. 1 (2013): 27. https://doi.org/10.1186/2041-1480-4-27

Stanković, Ranka, Cvetana Krstev, Nikola Vulović and Biljana Lazić. "Biblisha", 2014

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. "Rule-based Automatic Multi-word Term Extraction and Lemmatization". In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard et al. Paris, France: European Language Resources Association (ELRA), 2016

Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović and Olivera Kitanović. *Keyword-Based Search on Bilingual Digital Libraries*, 112–123. Cham: Springer International Publishing, 2017. http://dx.doi.org/10.1007/978-3-319-53640-8_10

Tsvetkov, Yulia and Shuly Wintner. "Extraction of Multi-word Expressions from Small Parallel Corpora". In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 1256–1264. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1944566.1944710

Vintar, Špela and Darja Fišer. "Harvesting Multi-Word Expressions from Parallel Corpora". In *Proceedings of the 6th International Conference on*

Language Resources and Evaluation (LREC 08), Marrakech, Morocco: European Language Resources Association (ELRA), 2008. http://www.lrec-conf.org/proceedings/lrec2008/

Šandrih, Branislava, Cvetana Krstev and Ranka Stanković. "Two Approaches to Compilation of Bilingual Multi-Word Terminology Lists from Lexical Resources". *Natural Language Engineering*, 2019

Xu, Yan, Luoxin Chen, Junsheng Wei, Sophia Ananiadou, Yubo Fan et al.. "Bilingual Term Alignment from Comparable Corpora in English Discharge Summary and Chinese Discharge Summary". *BMC bioinformatics* Vol. 16, no. 1 (2015): 149