Context Independent Term Mapper for European Languages

Mārcis Pinnis

University of Latvia
19 Raina Blvd., Riga, Latvia
marcis.pinnis@gmail.com

Tilde
Vienības gatve 75a, Riga, Latvia
marcis.pinnis@tilde.lv

Abstract

In this paper the author presents a new context independent method for bilingual term mapping using maximised character alignment maps. The method tries to particularly address mapping of multi-word terms and compound terms that are extracted from comparable corpora. The method allows integrating linguistic resources (e.g., probabilistic dictionaries and character based transliteration systems) that significantly increase the mapping recall while maintaining a stable precision. The term mapping method has been automatically evaluated using the *EuroVoc* thesaurus with varying availability of linguistic resources and on terms extracted from Latvian-English medical domain comparable corpus collected from the Web. The paper shows that the results significantly outperform previously reported results on the same evaluation corpus.

1 Introduction

Multi-lingual terminology is a valuable resource not only in human and machine translation (MT), but also in many other application domains, for instance, information retrieval, semantic analysis, question answering and others. Multi-lingual term glossaries can be automatically acquired from existing resources (monolingual lists of terms, parallel or comparable corpora, etc.) with the help of term mapping. Term mapping methods according to previous research in the field can be divided in two categories – context dependent methods and context independent methods.

The context dependent methods are applicable in situations when there is enough context from which to draw statistics. The necessary amount of context can differ depending on the methods. For instance, for term mapping in parallel data it can be enough to simply have one parallel document pair or a sentence-aligned parallel corpus (Federmann et al., 2012; Wolf et al., 2011; Lefever et al., 2009; Gaussier et al., 2000).

For under-resourced languages and numerous domains, however, parallel resources are scarce and not always available. Therefore, a more promising resource is comparable corpora, which has recently received much attention in the scientific community for its applicability in MT (Skadiņa et al., 2012). Most of the context-dependent methods designed for term mapping in comparable corpora, however, require relatively large corpora (e.g., hundreds or even thousands of documents) in order to calculate reliable cross-lingual association measures (Fung and Yee, 1998; Rapp, 1999; Shao and Ng, 2004; Morin and Daille, 2010). The proposed methods have also been focussed on language pairs with relatively simple morphology (e.g., German-English, French-English), but have not been thoroughly investigated for more complex languages (e.g., Finnish, Latvian, etc.). A recent study in the European Commission financed project TTC (2013) revealed that while the context-dependent methods by Morin et al. (2010) perform well for English-French, their applicability for English-Latvian is questionable because of a term mapping precision of below 5%. Laroche and Langlais (2010) also reported a relatively low precision (far below 50%) using context-dependent methods.

Context independent term mapping methods, however, are designed for situations when there is no context or the context is not large enough to draw statistics. Recent work on context independent term mapping has been done by Ștefănescu (2012) where a cognate similarity measure based on the Levenshtein distance (Levenshtein, 1966) was applied in order to estimate how similar two terms are. The method's weakness is a very limited term mapping recall.

Following previous work in context independent term mapping, this paper presents a new context independent term mapping method using

Proceedings of Recent Advances in Natural Language Processing, pages 562–570,
Hissar, Bulgaria, 7-13 September 2013.

maximised character alignment maps that has been created for term and term phrase mapping in term-tagged comparable corpora. The method allows mapping of multi-word terms and terms with different numbers of tokens in the source and target language parts – two term mapping scenarios that have not been sufficiently addressed by previous research. The mapper has been specifically designed to address term mapping between European languages (including languages with different alphabets based on Latin, Cyrillic and Greek) and it allows integrating linguistic resources to increase recall (while maintaining the same level of precision) of the mapped terms.

The mapper has been evaluated on the *EuroVoc* thesaurus (Steinberger et al., 2002) for 23 language pairs and for the Latvian-English language pair on a medical domain comparable corpus that was collected from the Web. The evaluation also shows benefits of having additional linguistic resources (e.g., probabilistic dictionaries, and transliteration support) with respect to having only some of the resources (or none at all) available.

The paper is structured so that section 2 describes the design of the term mapping system, section 3 describes the evaluation process and provides evaluation results with space constrained analysis, and the paper is concluded in section 4.

2 The Term Mapping Method

Given two lists of terms (in two different languages) the task of the term mapping system is to identify which terms from the source language contain translation equivalents in the target language. The system (as shown in Figure 1) consists of two main components – monolingual term pre-processing and term mapping. A possible third module that is not discussed in this paper is term pair consolidation – a language specific process that allows increasing term mapping precision by identifying morphological variability between term pairs and filtering out possible invalid mappings.

2.1 Term Pre-processing

Before mapping, all source and target language terms are tokenized and pre-processed using linguistic resources (if such are available). For each token the pre-processing module:

- Rewrites the token using lower-case letters;
- Rewrites the token with letters from the English alphabet (*simple transliteration*); letters that cannot be rewritten (e.g., the Russian softening and hardening marks "ъ" and "ь") are removed and letters that correspond to multiple letters in the English alphabet are expanded (e.g., the Russian "*ш*" and Latvian "*š*" are rewritten as "*sh*" in English).
- Finds top *N* translation equivalents in the other language using a probabilistic dictionary, e.g., in the *Giza++* format (Och and Ney, 2003).
- Finds top *M* transliteration equivalents in the other language using a *Moses* (Koehn et al., 2007) character-based SMT system.

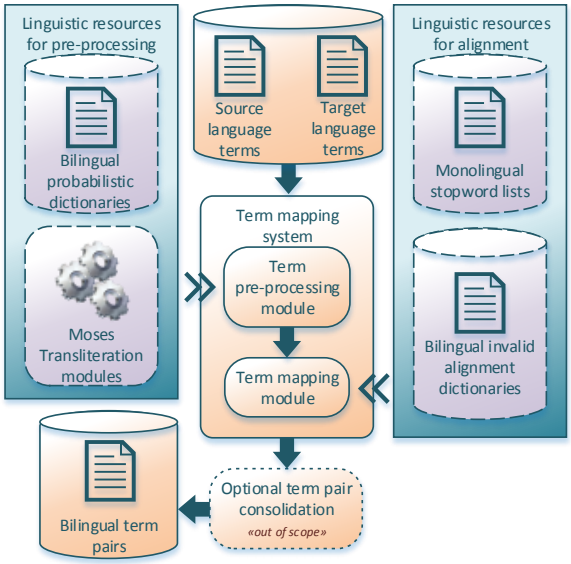

Figure 1: The overall design

Table 1 gives an example of a term in Latvian and English languages ("*extensive farming*") that has been pre-processed with direct *source-to-target* and *target-to-source* linguistic resources. If direct resources are not available, English can be used as an *Interlingua* for the dictionary-based look-up and the SMT-based transliteration.

The system allows limiting the retrieved candidates with confidence score thresholds, therefore, for the Latvian-to-English direction the example shows less than three transliteration candidates. For translation a limiting factor is also the available number of entries in the dictionary.

If for a language pair direct linguistic resources are not available, but there exist resources from the source and target languages to the English language, then the system allows using English as an *Interlingua* for term mapping.

Latvian term "*Ekstensīvā lauksaimniecība*"		
Lowercase form	ekstensīvā	lauksaimniecība
Simple translit.	ekstensiva	lauksaimnieciba
SMT translit.	extensiva extensive	-
Translation	-	agriculture farming
English term "*Extensive farming*"		
Lowercase form	extensive	farming
Simple translit.	extensive	farming
SMT translit.	ekstensīviem ekstensīvie ekstensīvai	farmēšana farmings farming
Translation	apjomīgam ekstensīvas izvērstāku	turēšanas saimniekošanas zemkopībā

Table 1: Examples of pre-processed terms

2.2 Term Mapping

After pre-processing the mapping module performs bi-directional term mapping. As shown in Figure 2 for each token in a term the mapping module operates with a set of constituents - *1* to *N* translation equivalents, *1* to *M* transliteration equivalents, one simple transliteration equivalent and one lowercased equivalent. The set of available constituents depends on the linguistic resources used (e.g., direct dictionaries, Interlingua dictionaries, no dictionaries, etc.).

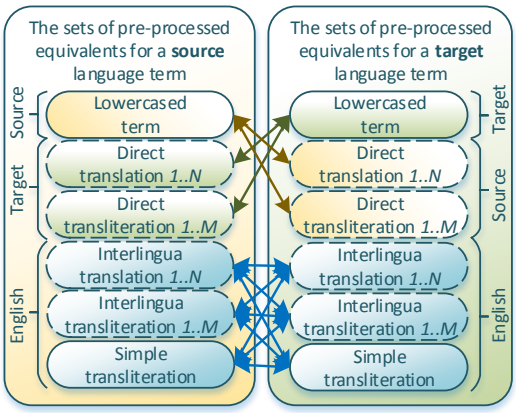

Figure 2: Bi-directional comparison sets for a single pre-processed term pair

The task of the mapping module is to decide whether a term pair can be mapped or not. The mapping process will be explained with the help of an example – the mapping of the English term "*dose of chemotherapy*" and its German translation "*chemotherapiedosis*". The mapping is performed in three steps.

2.2.1 Identification of Content Overlaps

At first, for every pre-processed token's constituent, we identify the *longest common substring* in all other term's pre-processed constituents that are in the same language (in Figure 2 comparison sets of the same language are connected with a bi-directional arrow). For the German-English example, the pre-processing module produced "*chemotherapiedosis*" as a simple transliteration of the German term. As the English lowercased term and the simple transliteration of the German term are within valid comparison sets, the mapper will analyse content overlaps between these constituents.

When identifying the *longest common substring* the positions of the substring within the constituents are preserved. If the length difference between the substring and the full source or target constituents exceeds a threshold (defined in a configuration file), the substring information is kept for the next step.

The results of the first step on the example are given in Figure 3. Two of the three English constituents ("*dose*" and "*chemotherapy*") can be nested within the German constituent. The third constituent's ("*of*") character overlap does not exceed the threshold (0.75 has been empirically selected as an appropriate default value), therefore, the substring information is ignored.

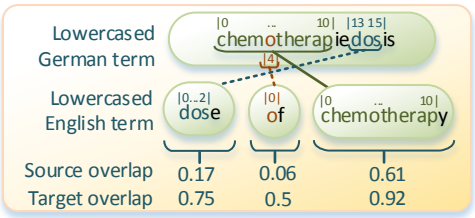

Figure 3: Longest common substring overlaps in German and English candidates

If the longest common substring overlap does not exceed the threshold, the mapper uses a fallback method based on the *Levenshtein distance* as applied by Ştefănescu (2012). The distance metric is transformed to a similarity metric:

$$Sim(s_1,s_2) = \frac{max(len(s_1),len(s_2))\text{-}LD(s_1,s_2)}{max(len(s_1),len(s_2))} \quad (1)$$

where *LD* is the *Levenshtein distance* between two strings, and *len* is a string length function. Each deletion, insertion and substitution is equal-

ly penalised with one point as in the first version of the *Levenshtein distance* (Levenshtein, 1966).

The motivation behind application of the alternative metric is that the SMT transliteration may introduce additional or different letters in a string and thus the longest common substring-based method can fail. However, this method has a limitation – it does not allow sub-word level mapping and if the similarity between two strings exceeds a predefined threshold, it is assumed that there is a complete overlap between the two strings. Assuming that the first comparison did not produce satisfactory results, Figure 4 shows the alternative comparison results for the example, however, none of the candidate pairs achieves a sufficient content overlap.

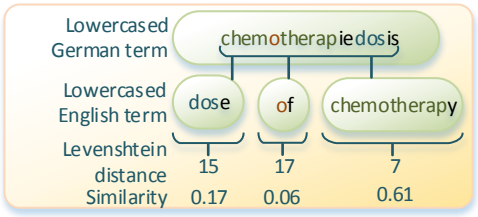

Figure 4: Levenshtein distance-based overlaps in German and English candidates

The result of this step is a list of binary alignment maps for constituent pairs. For instance, the binary alignment map for "*chemotherapiedosis*" and "*dose*" is "*0000000000011100*" (and "*1110*" for the target constituent).

2.2.2 Maximisation of content overlaps

In the next step the binary alignment lists are used to identify the mapping sequence that maximises the content overlap between the two terms. At first, the system iterates through the source term's tokens and tries to find for each token the constituent that has the highest overlap in a target term's constituent. At the same time the system maintains for each target term's token a binary one-dimensional alignment map that defines what part of the token has been already mapped in order not to allow conflicting and overlapping alignments. The length of the alignment map is determined by the longest constituent of the source and target terms. To find similar mappings from the target language, the iterative process is performed also for each token of the target term.

The example above contained two content overlaps (remember – the overlaps of the constituent "*of*" did not exceed thresholds). The overlap maximisation process in two iterations is shown in Figure 5.

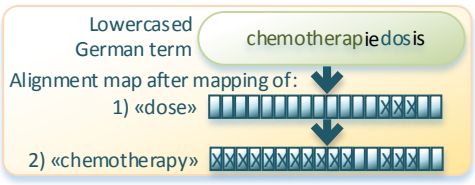

Figure 5: An example of the alignment map generation process for the German-English term pair

The goal of the mapper is to find term mappings that have a content overlap between terms in a way that restricts non-aligned segments (tokens or parts of tokens), but still allows a certain degree of imperfect mappings. For instance, we want the system to be able to decide that "*cost of treatment*" in English can be mapped to "*ārstēšanas izmaksas*" in Latvian (which is a direct translation) although it is evident that the token "*of*" does not have a mapping. However, we do not want the system to decide that "*β particles*" in English can be mapped to "*daļiņas*" in Latvian (transl. "*particles*") as well as we would not want "*electromagnetic field*" in English to be mapped to "*magnētiskais lauks*" in Latvian (transl. "*magnetic field*"). There is no perfect recipe that allows identifying all good and sufficient mappings from all bad and incomplete mappings in a language independent fashion, however, the mapper allows users to decide whether non-mapped segments at the beginning or the end of terms should be allowed or prohibited. Consequently the mapper can be executed in order to allow trimmed mappings, but not to limit non-mappings in-between of mapped segments. When trimmed mappings are allowed, it is important to disallow terms starting or ending with stopwords. Stopwords have shown to be very noisy in the probabilistic dictionaries (containing many false translations or context dependent translations). The mapper allows filtering out trimmed term mappings that start or end with stopwords if stopword lists are available.

2.2.3 Scoring of consolidated overlaps

In the final step the aligned constituents that produced character alignment map with the maximum content overlap are enrolled in two strings (source and target) in order to score the total overlap. The non-aligned source and target tokens (if there are any) are attached at the end of each string. At the same time, spaces are added

to the other string to simulate non-aligned tokens.

As both the probabilistic dictionaries and the SMT-based transliteration systems provide confidence scores for each candidate, these scores are used as negative multipliers to filter out term pairs that may potentially result in invalid mappings.

The enrolled strings are scored using the *Levenshtein distance*-based similarity metric (described in section 2.2.1) multiplied by the negative multipliers. In the example the *Levenshtein distance* between "*chemotherapydoseof*" (representing the English term) and "*chemotherapiedosis$$*" (representing the German term; "*$$*" represent two space symbols) is *6*; the *Levenshtein distance*-based similarity is *0.7*. The simple transliteration does not have a negative multiplier, therefore, the term pair is considered to be mapped if the 0.7 is higher than a threshold.

2.3 How to Acquire Linguistic Resources?

The mapper is able to use four types of optional linguistic resources (probabilistic dictionaries, external *Moses* SMT-based transliteration modules, invalid mapping dictionaries, and stopword lists).

The resources integrated in the mapper have been built using *Giza++* probabilistic dictionaries extracted from the *DGT-TM* parallel corpus (Steinberger, 2012):

- The dictionaries have been filtered by removing translation entries below a certain threshold and entries that contain symbols that are not allowed in the source and target language alphabets (out-of-the-box support is provided for all official European languages).
- Dictionary entries with the *Levenshtein distance*-based similarity measure higher than a threshold are assumed to be transliterations. These entries are used as the training data for the character-based *Moses* transliteration module. The mapper has out-of-the-box support for transliteration of terms in 22 languages (see automatic evaluation) into English (and vice versa).
- Word pairs that have a high *Levenshtein distance*-based similarity, but are not defined as translation entries within the dictionary (i.e., the index of the line where the words are found in the dictionary differs), are extracted for the invalid mapping dictionary. For instance, "*pants*" in English has a similarity measure of 1.0 with "*pants*" in Latvian (transl. as "*article*" or "*paragraph*"). The invalid mapping dictionary is used to filter possible invalid source and target token pairs before the first step of the mapping module.

3 Evaluation

The mapper has been evaluated using two evaluation methods – automated evaluation and manual evaluation. The automated evaluation was performed for language pairs included in the *EuroVoc* thesaurus. It shows the applicability of the method for European languages and allows estimation of the upper level of recall that can be expected on comparable Web corpora.
The manual evaluation was performed on terms mapped in a Latvian-English comparable Web corpora in the medical domain. This evaluation allows estimating the expected performance of the method in terms of precision on noisy data.

3.1 Automatic Evaluation

The automatic evaluation has three goals: 1) to show how additional linguistic resources influence term mapping, 2) to evaluate the performance on European language pairs, and 3) to compare results with previous research using the same evaluation corpus. The *EuroVoc* thesaurus was selected as a suitable test corpus for the automated evaluation because it covers 24 European languages, it contains a relatively large number of terms (at the time of evaluation – 6,797 terms for all languages except Hungarian with 6,790, Italian with 6,643, and Maltese with 987 terms), and in average 65.5% of terms across all languages are multi-word terms.

For each evaluated language pair two monolingual lists of terms were created. Because the mapper sees only two independent lists of terms, the search space for mapping is not 6,797 term pairs, but rather 46.2 million term pairs (e.g., 6,797*6,797 for English-Latvian). In this evaluation the highest matching target term is retrieved for each source term. For the language pairs for which additional resources are available, for every token a maximum of five transliterations and 10 dictionary translations are retrieved.

At first, the mapping performance when using direct (*source-to-target* and *target-to-source*) linguistic resources, Interlingua-based (*source-to-English* and *target-to-English*) resources, and no resources was analysed. Figure 6 shows results (in terms of precision "*P*" and recall "*R*") for the Latvian-Lithuanian language pair. It is evident that direct resources allow achieving sig-

nificantly higher recall than having Interlingua or no resources.

The results also suggest that the precision is stable at higher thresholds, however, it drops faster when using Interlingua-based resources. This can be explained by the noise that is introduced by the Interlingua-based resources. E.g., the term "*plakne*" (a type of a geometric figure) in Latvian can be wrongly be mapped to "*самолёт*" (a type of an aircraft) because both translate into English as "*plane*".

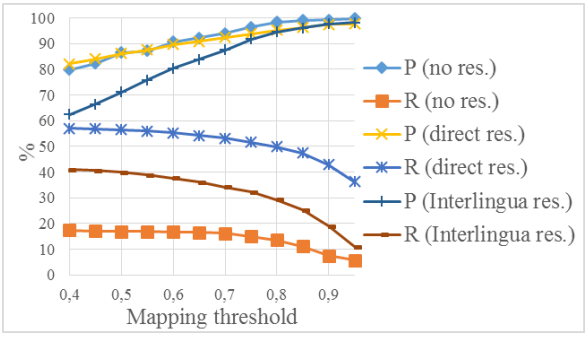

Figure 6: Latvian-Lithuanian evaluation results using direct, Interlingua, and no resources

Further, the benefits of having the probabilistic dictionaries and SMT-based transliteration modules were analysed. Figure 7 gives evaluation results for the Latvian-English language pair. The results show that without linguistic resources the recall is limited. This is due to the small number of terms that can be transliterated with the *simple transliteration* method. An analysis of 100 randomly selected unigram term pairs from the *EuroVoc* thesaurus revealed that 57 pairs were transliterations. 47 out of the 57 pairs were mapped using the *character-based transliteration* module. However, only 24 out of the 57 pairs were mapped using the *simple transliteration* method.

Evidently, adding resources allows significantly increasing the mapped term amount. It is also visible that the best results are achieved by using all linguistic resources.

Finally, term mapping was performed for 22 language pairs of the *EuroVoc* thesaurus with English as the source language. The results are given in Table 2. The evaluation was performed using direct *source-to-target* and *target-to-source* linguistic resources. The resources were built using *Giza++* probabilistic dictionaries extracted from the *DGT-TM* parallel corpus (Steinberger et al., 2012).

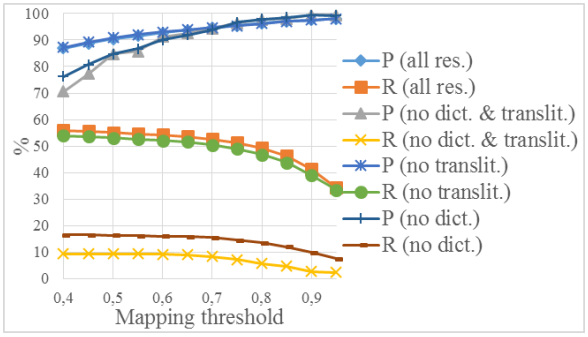

Figure 7: Latvian-English evaluation results using various resource configurations

The evaluation results show that the author's method significantly outperforms results reported earlier by Ştefănescu (2012) – an F1 score of 46.3 and 51.1 for English-Latvian and English-Romanian when using the same probabilistic dictionaries. The term mapping method proposed by Ştefănescu (2012) differs from the author's method in that it maps terms either with the *Levenshtein distance* based similarity metric or dictionary based exact match look-up. The author's proposed method, however, maps term tokens in sub-word level using maximised character alignment maps and applies Levenshtein distance just as a fall-back method and for scoring of the mapped term pairs.

Lang. pair	P	R	F1	Lang. pair	P	R	F1
en-mt	83.4	71.5	77.0	en-cs	85.9	53.4	65.8
en-fr	90.2	66.6	76.6	en-lt	86.1	52.6	65.3
en-ro	89.3	64.4	74.8	en-pl	86.0	52.1	64.9
en-es	91.1	63.2	74.6	en-el	86.0	49.6	62.9
en-pt	88.7	61.9	72.9	en-nl	82.0	50.7	62.7
en-it	87.4	62.0	72.6	en-sv	81.6	46.6	59.3
en-sk	90.8	58.8	71.4	en-da	81.4	45.3	58.2
en-lv	88.5	57.5	69.7	en-hu	78.5	45.7	57.8
en-sl	88.4	55.9	68.5	en-de	78.1	41.9	54.5
en-bg	88.0	55.2	67.9	en-et	74.5	39.0	51.2
en-hr	87.5	53.6	66.5	en-fi	72.3	33.7	46.0

Table 2: Evaluation results for *EuroVoc* language pairs with English as the source language (languages are given in the ISO 639-1 format).

The results suggest that the highest performance is achieved for the English-Maltese language pair, however, it is not comparable to the remaining results as they are based on only 987 term pairs from the *EuroVoc* thesaurus (covering mostly location and organisation named entities, which explains the relatively high recall).

An important aspect taken into account when designing the mapper was the mapping speed.

For the evaluation in Table 2 the mapper required in average 86.8 minutes (which is a speed of 8,868 term pairs per second) for one language pair on an 8 thread (4 core) Windows machine. The speed can be significantly improved by limiting the number of translation and transliteration candidates retrieved from the probabilistic dictionary and the character-based SMT module. The mapper requires in average less than 7 minutes for a language pair if no linguistic resources are used.

3.2 Manual Evaluation

The automatic evaluation was performed using terms in their base forms. The manual evaluation, therefore, has three goals: 1) to show the methods applicability on Web crawled comparable corpora 2) to show the methods performance in under-resourced conditions (the medical domain is out-of-domain for the DGT-TM corpus), 3) to show that the method can be applied for morphologically rich languages. The manual evaluation was performed for the Latvian-English language pair and for terms in the medical domain. Latvian was selected as one of the languages for this evaluation as it is a morphologically rich language and it is important to show that the method can be easily applicable to languages where terms are not always in their base forms.

Following the term mapping workflow proposed by Pinnis et al. (2012), two monolingual corpora were collected from the Web using the *Focussed Monolingual Crawler* (Mastropavlos and Papavassiliou, 2011). The acquired corpora (12,697 Latvian and 21,900 English documents) were then aligned in document level with the *DictMetric* (Su and Babych, 2012) comparability metric (59,600 document pairs were produced). The terms were tagged in the monolingual documents with *TWSC* (Pinnis et al., 2012). The term tagging step produced a total of 198,401 unique Latvian and 352,934 unique English terms. The reason why document alignment is a necessary step before mapping can be easily explained with the large number of monolingual terms. If the terms would be mapped between the two monolingual lists, the mapper would have to handle a search space of 70 billion term pairs and require over 91 days to complete (using direct linguistic resources). With document alignments the required time can be reduced to less than 2 days.

Finally, terms were bilingually mapped in the 59,600 document pairs. A maximum of three transliteration and translation candidates were retrieved for each token of a term. A total of 24,804 term pairs were produced above a threshold of 0.6 (for each source term only the target language term with the highest confidence score was returned). 1000 randomly selected term pairs were manually evaluated and the results are given in Table 3. The results are also compared with the method proposed by Ştefănescu (2012) using the same probabilistic dictionary.

The results suggest that the author's method performs significantly better for multi-word term mapping, which is the main goal of this method. It is also evident that the majority of true positives are scored with a mapping score of over 0.8. The results, however, require deeper analysis of why the unigram mapping score of the proposed method drops so fast.

Threshold	All terms		Multi-word terms		Single-word terms	
	Pairs	P	Pairs	P	Pairs	P
Author' s method (random 1000/24,804 term pairs):						
1.0	17	88.2%	0	-	17	88.2%
0.9	601	91.3%	111	85.6%	490	92.7%
0.8	724	85.6%	160	73.8%	564	89.0%
0.7	880	74.8%	203	65.0%	677	77.7%
0.6	1000	66.6%	267	50.6%	733	72.4%
Ştefănescu (2012) (random 1000/2,330 term pairs):						
1.0	25	84.0%	2	0.0%	23	91.3%
0.9	44	90.9%	7	71.4%	37	94.6%
0.8	88	93.2%	12	83.3%	76	94.7%
0.7	186	87.6%	46	65.2%	140	95.0%
0.6	387	73.6%	173	49.7%	214	93.0%
0.5	1000	44.8%	697	25.1%	303	90.1%

Table 3: Manual evaluation results on the medical domain Latvian-English comparable corpus

Another important question left to answer is whether the mapper finds term pairs that are unknown to the linguistic resources integrated in the mapper. The mapping method is only useful if it is able to identify *out-of-vocabulary* (OOV) term pairs. Therefore, the 1000 randomly selected term pairs from the manual evaluation were looked up in the probabilistic dictionary (for the 733 single-word terms) and in a translation model of an SMT system (for the 267 multi-word terms) that was trained on the same parallel corpus from which the probabilistic dictionary was created. The results of the analysis in comparison with the method proposed by Ştefănescu (2012) are given in Table 4.

Table 4 shows that 76.3% of all multi-word term pairs, which were evaluated as "*correct*" during the manual evaluation, could not be found

in the translation model of the SMT system. The results also suggest that the probabilistic dictionary introduces mapping errors as 24.75% of the wrongly mapped single-word term pairs were present in the dictionary.

	Single-word term pairs in the probabilistic dictionary		Multi-word term pairs in the Moses phrase table	
Evaluation:	**Correct**	**Wrong**	**Correct**	**Wrong**
Author's method:				
Source term OOV rate	13.94%	75.25%	76.30%	97.73%
Target term OOV rate	14.50%	75.66%	75.19%	97.73%
Term pair OOV rate	13.94%	75.25%	76.30%	97.73%
Ştefănescu (2012):				
Source term OOV rate	9.72%	76.00%	63.58%	99.58%
Target term OOV rate	12.09%	80.00%	62.86%	99.62%
Term pair OOV rate	12.09%	80.00%	62.86%	99.62%

Table 4: OOV analysis of randomly selected Latvian-English term pairs

4 Conclusion and Future Work

In this paper the author presented a new bilingual term mapping method using maximised character alignment maps. The method has been designed to address multi-word term pair as well as compound term pair mapping for European Languages that are based on Latin, Greek and Cyrillic alphabets.

The method has been automatically evaluated using the *EuroVoc* thesaurus for 23 language pairs. The paper discussed the impact of different linguistic resources on the term mapping performance. The method was also manually evaluated on terms mapped in a comparable corpus in the medical domain for the Latvian-English language pair, showing that the mapping method is suitable for handling noisy data collected from the Web. The evaluation also shows that up to 76.3% of the correctly mapped multi-word term pairs are out-of-vocabulary term pairs. The proposed term mapping method is able to find multi-word term alignments with a relatively high precision of up to 85.6%. It should, however, be noted that the scores depend on the corpus processed and may differ between language pairs as seen in the automatic evaluation.

The term mapping toolkit together with configuration and evaluation recipes is released under a non-commercial (free to use for scientific purposes) license. The toolkit can be downloaded from https://github.com/pmarcis/mp-aligner. The linguistic resources for the above-mentioned language pairs are also included in the release.

The future work on the term mapping method will involve a more in-depth error analysis of the mapped term pairs. Preliminary analysis suggests that simple filtering techniques could be applied to increase precision even further. For comparable corpora evaluation scenarios comparison with context-dependent methods is also necessary. The application of machine learning methods needs to be investigated in order to fine-tune the system's parameters for specific language pairs in order to achieve higher recall and precision. As the produced bilingual term pairs can be beneficial for MT systems, it is also necessary to evaluate the applicability of the method for MT system adaptation purposes to narrow domains. An important future step in order to improve the precision of term mapping and in order to provide term pairs for automated integration into terminology data bases in bilingual term extraction (of which term mapping is an integral component) is also term pair consolidation with knowledge rich term normalisation methods or language independent statistical methods that require presence of a large reference corpus.

Acknowledgement

This work has been supported by the European Social Fund within the project «*Support for Doctoral Studies at University of Latvia*». The research within the project TaaS leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), Grant Agreement no 296312.

References

Federmann, C., Gromann, D., Declerck, T., Hunsicker, S., Krieger, H., & Budin, G. (2012). Multilingual Terminology Acquisition for Ontology-based Information Extraction. In Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012) (pp. 166–175). Madrid, Spain.

Fung, P., & Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (pp. 414–420). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gaussier, E., Hull, D. A., Salah, A., & Ait-Mokhtar, S. (2000). Term Alignment in Use: Machine-Aided Human Translation. Véronis, Jean: Parallel Text Processing. Alignment and Use of Translation Corpora. Dordrecht, 253–274.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Laroche, A., & Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 617–625). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lefever, E., Macken, L., & Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 496–504). Stroudsburg, PA, USA: Association for Computational Linguistics.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10: 707–10.

Mastropavlos, N., & Papavassiliou, V. (2011). Automatic acquisition of bilingual language resources. In Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece.

Morin, E., & Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. Language Resources and Evaluation, 44, 79–95.

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2010). Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. ACM Transactions on Speech and Language Processing (TSLP), 7(1), 1.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29, 19–51.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012) (pp. 193–208). Madrid.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 519–526). Stroudsburg, PA, USA: Association for Computational Linguistics.

Shao, L., & Ng, H. T. (2004). Mining new word translations from comparable corpora. In Proceedings of the 20th international conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1220355.1220444

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M.L., & Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (pp. 438–445). Istanbul, Turkey: European Language Resources Association (ELRA).

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlter, P. (2012). Dgt-tm: A freely available translation memory in 22 languages. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (pp. 454–459).

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. Computational Linguistics and Intelligent Text Processing, 101–121.

Su, F., & Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (pp. 10–19). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In The 5th Workshop on Building and Using Comparable Corpora (pp. 98–103).

TTC Project. (2013). Public deliverable D7.3: Evaluation of the impact of TTC on Statistical MT (p. 38). TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora. Retrieved from http://ttc-project.eu/images/stories/TTC_D7.3.pdf

Wolf, P., Bernardi, U., Federmann, C., & Hunsicker, S. (2011). From Statistical Term Extraction to Hybrid Machine Translation. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation (pp. 379–419).