Cross-lingual WSD for Translation Extraction from Comparable Corpora

Marianna Apidianaki
LIMSI-CNRS
Rue John Von Neumann
BP 133, 91403
Orsay Cedex, France
marianna@limsi.fr

Nikola Ljubešić
Dept. of Information Sciences
University of Zagreb
Ivana Lučića 3, HR-10000
Zagreb, Croatia
nljubesi@ffzg.hr

Darja Fišer
Department of Translation
University of Ljubljana
Aškerčeva 2, SI-1000
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Abstract

We propose a data-driven approach to enhance translation extraction from comparable corpora. Instead of resorting to an external dictionary, we translate source vector features by using a cross-lingual Word Sense Disambiguation method. The candidate senses for a feature correspond to sense clusters of its translations in a parallel corpus and the context used for disambiguation consists of the vector that contains the feature. The translations found in the disambiguation output convey the sense of the features in the source vector, while the use of translation clusters permits to expand their translation with several variants. As a consequence, the translated vectors are less noisy and richer, and allow for the extraction of higher quality lexicons compared to simpler methods.

1 Introduction

Large-scale comparable corpora are available in many language pairs and are viewed as a source of valuable information for multilingual applications. Identifying translation correspondences in this type of corpora permits to construct bilingual lexicons for low-resourced languages, and to complement and reduce the sparseness of existing resources (Munteanu and Marcu, 2005; Snover et al., 2008). The main assumption behind translation extraction from comparable corpora is that a source word and its translation appear in similar contexts (Fung, 1998; Rapp, 1999). So, in order to identify a translation correspondence between the two languages, the contexts of the source word and the candidate translation have to be compared. For this comparison to take place, the same vector space has to be produced, which means that the vectors of the one language have to be translated in the other language. This generally assumes the availability of a bilingual dictionary which might however not be the case for some language pairs and domains. Moreover, the classic way in which a dictionary is put into use, which consists in translating vector features by their first translation in the dictionary, neglects semantics. We expect that a method capable of identifying the correct sense of the features and translating them accordingly could contribute to producing cleaner vectors and to extracting higher quality lexicons.

In this paper, we show how source vectors can be translated into the target language by a cross-lingual Word Sense Disambiguation (WSD) method which exploits the output of data-driven Word Sense Induction (WSI) (Apidianaki, 2009), and demonstrate how feature disambiguation enhances the quality of the translations extracted from the comparable corpus. This study extends our previous work on the topic (Apidianaki et al., 2012) by applying the proposed methods to a comparable corpus of general language (built from Wikipedia) and optimizing various parameters that affect the quality of the extracted translations. We expect the disambiguation to have a beneficial impact on the results given that polysemy is a frequent phenomenon in a general, mixed-domain corpus. Our experiments are carried out on the English-Slovene language pair but as the methods are totally data-driven, the approach can be easily applied to other languages.

The paper is organized as follows: In the next section, we present some related work on bilingual lexicon extraction from comparable corpora. Section 3 presents the data used in our experiments and Section 4 provides details on the approach and the experimental setup. In Section 5, we report and discuss the obtained results before concluding and presenting some directions for future work.

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 1–10,
Sofia, Bulgaria, August 8, 2013.

2 Related work

The traditional approch to translation extraction from comparable corpora and most of its extensions (Fung, 1998; Rapp, 1999; Shao and Ng, 2004; Otero, 2007; Yu and Tsujii, 2009; Marsi and Krahmer, 2010) presuppose the availability of a bilingual lexicon for translating source vectors into the target language. A translation candidate is generally considered as correct if it is an appropriate translation for at least one sense of the source word in the dictionary, which often corresponds to its most frequent sense. An alternative consists in considering all translations provided for a word in the dictionary but weighting them by their frequency in the target language (Prochasson et al., 2009; Hazem and Morin, 2012). The high quality of the exploited handcrafted resources, combined to the skewed distribution of the translations corresponding to different word senses, often lead to satisfying results. Nevertheless, the applicability of the methods is limited to languages and domains where bilingual resources are available. Moreover, by promoting the most frequent sense/translation, this approach neglects polysemy. We believe that feature disambiguation can lead to the production of cleaner vectors and, consequently, to higher quality results.

The need to bypass pre-existing dictionaries has been addressed by Koehn and Knight (2002) who built the initial seed dictionary automatically, based on identical spelling features between English and German. Cognate detection has also been used by Saralegi et al. (2008) for extracting word translations from English-Basque comparable corpora. The cognate and seed lexicon approaches have been successfully combined by Fišer and Ljubešić (2011) who showed that the results with an automatically created seed lexicon, based on language similarity, can be as good as with a pre-existing dictionary. But all these approaches work on closely-related languages and cannot be used as successfully for language pairs with little lexical overlap, such as English and Slovene, which is the case in this experiment.

Regarding the translation of the source vectors, we use contextual information to disambiguate their features and translate them using clusters of semantically similar translations in the target language. A similar idea has been implemented by Kaji (2003) who performed sense-based word clustering to extract sets of synonymous translations from comparable corpora with the help of a bilingual dictionary.

Using translation clusters permits to expand feature translation and to suggest multiple semantically correct translations. A similar approach has been adopted by Déjean et al. (2005) who expand vector translation by using a bilingual thesaurus instead of a lexicon. In contrast to their work, the method proposed here does not rely on any external knowledge source to determine word senses or translation equivalents, and is thus fully data-driven and language independent.

3 Resources

3.1 Comparable corpus

The comparable corpus from which the bilingual lexicon will be extracted is a collection of English (EN) and Slovene (SL) texts extracted from Wikipedia. The February 2013 dumps of Wikipedia articles were downloaded and cleaned for both languages after which the English corpus was tokenized, part-of-speech (PoS) tagged and lemmatized with the TreeTagger (Schmid, 1994). The same pre-processing was applied to the Slovene corpus with the ToTaLe analyzer (Erjavec et al., 2010) which uses the TnT tagger (Brants, 2000) and was trained on MultextEast corpora. The Wikipedia corpus contains about 1.5 billion tokens for English and almost 24 million tokens for Slovene.

In previous work, we applied our approach to a specialized comparable corpus from the health domain (Apidianaki et al., 2012). The results were encouraging, showing how translation clustering and vector disambiguation help to improve the quality of the translations extracted from the comparable corpus. We believe that the positive impact of this approach will be more significant on lexicon extraction from a general language comparable corpus, in which polysemy is more prominent.

3.2 Parallel corpus

The parallel corpus used for clustering and word sense induction consists of the Slovene-English parts of Europarl (release v6) (Koehn, 2005) and of JRC-Acquis (Steinberger et al., 2006) and amounts to approximately 35M words per language. A number of pre-processing steps are applied to the corpus prior to sense induction, such

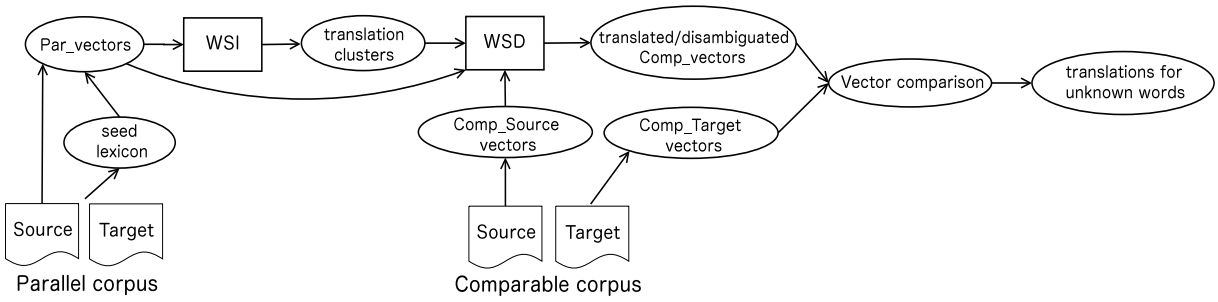

Figure 1: Translation extraction from comparable corpora using cross-lingual WSI and WSD.

as elimination of sentence pairs with a great difference in length, lemmatization and PoS tagging with the TreeTagger (for English) and ToTaLe (for Slovene) (Erjavec et al., 2010). Next, the corpus is word-aligned with GIZA++ (Och and Ney, 2003) and two bilingual lexicons are extracted, one for each translation direction (EN–SL/SL–EN). To clean the lexicons from noisy alignments, the translations are filtered on the basis of their alignment score and PoS, keeping only translations that pertain to the same grammatical category as the source word. We retain only intersecting alignments and use for clustering translations that translate a source word more than 10 times in the training corpus. This threshold reduces data sparseness issues that affect the clustering and eliminates erroneous word alignments. The filtered EN-SL lexicon contains entries for 6,384 nouns, 2,447 adjectives and 1,814 verbs having more than three translations in the training corpus.

The parallel corpus, which contains EU texts, is more specialized than the comparable corpus built from Wikipedia. This is not the ideal scenario for this experiment; domain adaptation is important for the type of semantic processing we want to apply as there might be a shift in the senses present in the two corpora. However, as EU texts often contain a lot of general vocabulary, we expect that this discrepancy will not strongly affect the quality of the results.

3.3 Gold standard

We evaluate the quality of the bilingual lexicons extracted from the comparable corpus by comparing them to a gold standard lexicon, which was built from the aligned English (Fellbaum, 1998) and Slovene wordnets (Fišer and Sagot, 2008). We extracted all English synsets from the Base Concept sets that belong to the Factotum domain and contain literals with polysemy levels 1-5 and their Slovene equivalents which have been validated by a lexicographer. Of 1,589 such synsets, 200 were randomly selected and used as a gold standard for automatic evaluation of the method proposed in this paper.

4 Experimental setup

4.1 Overview of the method

Figure 1 gives an overview of the way information mined from the parallel training corpus is exploited for discovering translations of source (English) words in the comparable corpus. The parallel corpus serves to extract an English-Slovene seed lexicon and source language context vectors (Par_vectors) for the Slovene translations of English words. These vectors form the input to the Word Sense Induction (WSI) method which groups the translations of an English word into clusters.

The clusters of semantically related Slovene translations constitute the candidate senses which, together with the Par_vectors, are used for disambiguating and translating the vectors extracted from the source (English) side of the comparable corpus (Comp_source). The translated vectors are then compared to the ones extracted from the target language (Slovene) side of the comparable corpus (Comp_target) and the best translations are selected, for a list of unknown words. All steps of the proposed method illustrated in Figure 1 will be detailed in the following sections.

4.2 Translation clustering

The translations of the English words in the lexicon built as described in 3.2 are clustered according to their semantic proximity using a cross-lingual Word Sense Induction method (Apidianaki, 2008). For each translation T_i of a word w, a vector is built from the content word co-

Language	POS	Source word	Slovene sense clusters
EN–SL	Nouns	sphere	{krogla}_(*geometrical shape*) {sfera, področje}_(*area*)
		address	{obravnava, reševanje, obravnavanje}_(*dealing with*) {naslov}_(*postal address*)
		portion	{kos}_(*piece*) {obrok, porcija}_(*serving*) {delež}_(*share*)
		figure	{številka, podatek, znesek}_(*amount*) {slika}_(*image*) {osebnost}_(*person*)
	Verbs	seal	{tesniti}_(*to be water-/airtight*) {zapreti, zapečatiti}_(*to close an envelope or some other container*)
		weigh	{pretehtati}_(*consider possibilities*) {tehtati, stehtati}_(*check weight*)
		educate	{poučiti}_(*give information*) {izobraževati, izobraziti}_(*give education*)
		consume	{potrošiti}_(*spend money/goods*) {uživati, zaužiti}_(*eat/drink*)
	Adjs	mature	{zrel, odrasel}_(*adult*) {zorjen, zrel}_(*ripe*)
		minor	{nepomemben}_(*not very important*) {mladoleten, majhen}_(*under 18 years old*)
		juvenile	{nedorasel}_(*not adult/biologically mature yet*) {mladoleten, mladoletniški}_(*not 18/legally adult yet*)
		remote	{odmaknjen, odročen}_(*far away and not easily accessible*) {oddaljen daljinski}_(*controlled from a distance (e.g. remote control)*)

Table 1: Entries from the English-Slovene sense cluster inventory.

occurrences of w in the parallel sentences where it is translated by T_i. Let N be the number of features retained for each T_i from the corresponding source contexts. Each feature F_j ($1 \leq j \leq N$) receives a total weight with a translation T_i, $\mathrm{tw}(F_j, T_i)$, defined as the product of the feature's global weight, $\mathrm{gw}(F_j)$, and its local weight with that translation, $\mathrm{lw}(F_j, T_i)$. The global weight of a feature F_j is a function of the number N_i of translations (T_i's) to which F_j is related, and of the probabilities (p_{ij}) that F_j co-occurs with instances of w translated by each of the T_i's:

$$\mathrm{gw}(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{ij})}{N_i} \quad (1)$$

Each p_{ij} is computed as the ratio of the co-occurrence frequency of F_j with w when translated as T_i to the total number of features seen with T_i:

$$p_{ij} = \frac{\mathrm{cooc\_frequency}(F_j, T_i)}{N} \quad (2)$$

The local weight $\mathrm{lw}(F_j, T_i)$ between F_j and T_i directly depends on their co-occurrence frequency:

$$\mathrm{lw}(F_j, T_i) = \log(\mathrm{cooc\_frequency}(F_j, T_i)) \quad (3)$$

The pairwise similarity of the translations is calculated using the Weighted Jaccard Coefficient (Grefenstette, 1994).

$$WJ(T_m, T_n) = \frac{\sum_j \min(tw(T_m, F_j), tw(T_n, F_j))}{\sum_j \max(tw(T_m, F_j), tw(T_n, F_j))} \quad (4)$$

The similarity score of each translation pair is compared to a threshold locally defined for each w using an iterative procedure. The threshold (T) for a word w is initially set to the mean of the scores (above 0) of its translation pairs. The set of translation pairs of w is then divided into two sets ($G1$ and $G2$) according to whether they exceed, or are inferior to, the threshold. The average of scores of the translation pairs in each set is computed ($m1$ and $m2$) and a new threshold is calculated that is the average of $m1$ and $m2$ ($T = (m1 + m2)/2$). The new threshold serves to separate again the translation pairs into two sets, a new threshold is calculated and the procedure is repeated until convergence.

The semantically similar translations of w are grouped into clusters. Translation pairs with a score above the threshold form initial clusters that

might be further enriched provided that there exist additional strongly related translations. Clustering stops when all translations of *w* are clustered and all their relations have been checked. An important feature of the algorithm is that it performs soft clustering, so translations can be found in different clusters. The final clusters are characterized by global connectivity, i.e. all their elements are linked by pertinent relations.

Table 1 gives examples of clusters obtained for English words of different PoS with clear sense distinctions in the parallel corpus. For each English word, we provide the obtained clusters of Slovene translations including a description of the sense described by each cluster. For instance, the translations for the adjective *minor* from the training corpus (*nepomemben*, *mladoleten* and *majhen*) are grouped into two clusters describing its two senses: {nepomemben} - "not very important" and {mladoleten, majhen} - "under 18 years old". The resulting cluster inventory contains 13,352 clusters in total, for 8,892 words. 2,585 of the words (1,518 nouns, 554 verbs and 513 adjectives) have more than one cluster.

In the next section, we explain how the clusters and the corresponding translation vectors are used for disambiguating the source language vectors extracted from the comparable corpus.

4.3 Cross-lingual vector comparison

4.3.1 Vector building

We build context vectors in the two languages for nouns occurring at least 50 times in the comparable corpus. The frequency threshold is important for the lexicon extraction approach to produce good results. As features we use three content words to the left and to the right of the retained nouns, stopping at the sentence boundary, without taking into account their position. Log-likelihood is used to calculate feature weights.

In the reported experiments we focus on the 1,000 strongest features. A portion of these features is disambiguated for each headword, depending on the availability of clustering information. We observed that disambiguating a smaller amount of features yielded similar results and including additional features did not improve the results.

4.3.2 Vector translation and disambiguation

Translation correspondences between the two languages of the comparable corpus are identified by comparing the source language vectors, built as described in Section 4.3.1, to the ones of the candidate translations. This comparison serves to quantify the similarity of the source and target words represented by the vectors and the highest ranked pairs are retained.

For the comparison to take place, the source vectors have to be translated in the target language. In most previous work, the vectors were translated using external seed dictionaries: the first translation proposed for a word in the dictionary was used to translate all instances of the word in the vectors irrespective of their sense. Here, we replace the external dictionary with the output of a data-driven cross-lingual WSD method (Apidianaki, 2009) which renders the method knowledge light and adaptable to other language pairs.

The translation clusters obtained during WSI (cf. Section 4.2) describe the senses of the English words in the parallel corpus. We exploit this sense inventory for disambiguating the features in the English vectors extracted from the comparable corpus. More precisely, we ask the WSD method to select among the available clusters the one that correctly translates in Slovene the sense of the English features in the vectors built from the comparable corpus. The selection is performed by comparing information from the context of a feature, which corresponds to the rest of the vector where the feature appears, to the source language vectors of the translations which served to their clustering. Inside the vectors, the features are ordered according to their score, calculated as described in Section 4.3.1. Feature weights filter out the *weak* features, i.e. features with a score below the experimentally set threshold of 0.01. The retained features are then considered as a bag of words.

On the clusters' side, the information used for disambiguation is found in the source language vectors that revealed the similarity of the translations. If common features (CFs) exist between the context of a feature and the vectors of the translations in a cluster, a score is calculated corresponding to the mean of the weights of the CFs with the clustered translations, where weights correspond to the total weights (*tw*'s) computed between features and translations during WSI. In formula 5, CF_j is the set of CFs and N_{CF} is the number of translations T_i characterized by a CF.

$$wsd\_score = \frac{\sum_{i=1}^{N_{CF}} \sum_j w(T_i, CF_j)}{N_{CF} \cdot |CF_j|} \qquad (5)$$

PoS	Feature	Assigned Cluster	MFT
Nouns	party	{oseba, stran, pogodbenica, stranka}	stranka
	matter	{zadeva, vprašanje}	zadeva
Verbs	settle	{urediti, rešiti, reševati}	rešiti
	follow	{upoštevati, spremljati, slediti}	slediti
Adjs	alternative	{nadomesten, alternativen}	alternativen
	involved	{vključen, vpleten}	vključen

Table 2: Disambiguation results.

The cluster that receives the highest score is selected and assigned to the feature as a sense tag. The features are also tagged with their most frequent translation (MFT) in the parallel corpus, which sometimes already exists in the cluster selected during WSD.

In Table 2, we present examples of disambiguated features of different PoS from the vector of the word *transition*. The context used for disambiguation consists of the other strong features in the vector and the cluster that best describes the sense of the features in this context is selected. In the last column, we provide the MFT of the feature in the parallel corpus. In the examples shown here the MFT translation already exists in the cluster selected by the WSD method but this is not always the case. As we will show in the Evaluation section, the configuration where the MFT from the cluster assigned during disambiguation is selected (called CLMFT) gives better results than MFT, which shows that the MFT in the selected cluster is not always the most frequent alignment for the word in the parallel corpus. Furthermore, the clusters provide supplementary material (i.e. multiple semantically correct translations) for comparing the vectors in the target language and improving the baseline results. Still, MFT remains a very powerful heuristic due to the skewed distribution of word senses and translations.

4.4 Vector comparison

The translation clusters proposed during WSD for the features in the vectors built from the source side of the comparable corpus serve to translate the vectors in the target language. In our experiments, we compare three different ways of translating the source language features.

1. by keeping the most frequent translation/alignment of the feature in the parallel corpus (MFT);
2. by keeping the most frequent translation from the cluster assigned to the feature during disambiguation (CLMFT); and
3. by using the same cluster as in the second approach, but producing features for all translations in the cluster with the same weight (CL).

The first approach (MFT) serves as the baseline since, instead of the sense clustering and WSD results, it just uses the most frequent sense/alignment heuristic. In the first batch of experiments, we noticed that the results of the CL and CLMFT approaches heavily depend on the part-of-speech of the features. So, we divided the CL and CLMFT approaches into three sub-approaches:

1. translate only nouns, verbs or adjectives with the clusters and other features with the MFT approach (CLMFT_N, CLMFT_V, CLMFT_A);
2. translate nouns and adjectives with the clusters and verbs with the MFT approach (CLMFT_NA); and
3. translate nouns and verbs with the clusters and adjectives with the MFT approach (CLMFT_NV).

The distance between the translated source and the target-language vectors is computed by the Dice metric. By comparing the translated source vectors to the target language ones, we obtain a ranked list of candidate translations for each gold standard entry.

5 Evaluation

5.1 Metrics

The final result of our method consists in ranked lists of translation candidates for gold standard entries. We evaluate this output by the mean reciprocal rank (MRR) measure which takes into account

the rank of the first good translation found for each entry. Formally, MRR is defined as

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (6)$$

where $|Q|$ is the length of the query, i.e. the number of gold standard entries we compute translation candidates for, and $rank_i$ is the position of the first correct translation in the candidate list.

5.2 Results

Table 4 shows the translation extraction results for different configurations. The MFT score is used as the baseline. We observe that disambiguating all features in the vectors (CL) yields lower results than the baseline compared to selecting only the most frequent translation from the cluster which slightly outperforms the MFT baseline. In the CLMFT_N, CLMFT_NA, CLMFT_NV configurations we disambiguate noun features, nouns and adjectives, and nouns and verbs, respectively, and translate words of other PoS using the MFT. In CLMFT_N, for instance, nouns are disambiguated while verbs and adjectives are translated by the word to which they were most frequently aligned in the parallel corpus. The three configurations where nouns are disambiguated (CLMFT_N, CLMFT_NA, CLMFT_NV) give better results compared to those addressing verbs or adjectives alone. Interestingly, disambiguating only adjectives gives worse results than disambiguating only verbs, but the combination of nouns and adjectives outperforms the combination of nouns and verbs.

In CLMFT, features of all PoS are disambiguated but we only keep the most frequent translation in the cluster and ignore the other translations. This setting gives much better results than CL, where the whole cluster is used, which highlights two facts: first, that disambiguation is beneficial for translation extraction and, second, that the noise present in the automatically built clusters harms the quality of the translations extracted from the comparable corpus. The better score obtained for CLMFT compared to MFT also shows that, in many cases, the most frequent translation in the cluster does not coincide with the most frequent alignment of the word in the parallel corpus. So, disambiguation helps to select a more appropriate translation than the MFT approach. This improvement compared to the baseline shows again that WSD is useful in this setting.

	MRR
MFT	0.0685
CLMFT	0.0807
CL	0.0434
CLMFT_N	0.0817
CLMFT_A	0.07
CLMFT_V	0.0714
CLMFT_NA	0.0842
CLMFT_NV	0.08048

Table 3: Results of the experiment.

		MRR diff	p-value
MFT	CLMFT	0.0122	0.1830
MFT	CL	0.0251	0.0410
CLMFT	CL	0.0373	0.0120
MFT	CLMFT_NA	0.0157	0.4296
MFT	CLMFT_NV	0.0120	0.5195

Table 4: Comparison of different configurations.

In Table 4, the results for different configurations are compared. The statistical significance of the difference in the results was calculated by approximate randomization (1,000 repetitions). We observe that the differences between the CL and MFT configurations and the CL and CLMFT ones, are statistically significant. This confirms that taking most frequent translations, disambiguated or not, works better than exploiting all the information in the clusters. The remainder of the differences in the results are not statistically significant. One could wonder why the p-values are that high in case of the MFT setting on one side and CLMFT_NA and CLMFT_NV settings on the other side although the differences in the results are not that high. The most probable explanation is that there is a low intersection in correct results and errors. Because of that, flipping the results between the two systems – as performed in approximate randomization – often generates differences higher than the initial difference on the original results.

5.3 Qualitative analysis

Manual evaluation of the results shows that the procedure can deal with concrete words much better than with abstract ones. For example, the correct translation of the headword *enquiry* is the third highest-ranked translation. The results are

also much better with monosemous and domain-specific terms (e.g. the correct translation for *cataclysm* is the top-ranking candidate). On the other hand, general and polysemous expressions that can appear in a wide range of contexts are a much tougher nut to crack. For example, the correct translation candidate for word *role*, which can be used in a variety of contexts as well as metaphorically, is in the tenth position, whereas no correct translation was found for *transition*. However, it must be noted that even if the correct translation is not found in the results, the output of our method is in most cases a very coherent and solid description of the semantic field of the headword in question. This means that the list can still be useful for lexicographers to illicit the correct translation that is missing, or organize the vocabulary in terms of their relational-semantic principles.

We have also performed an error analysis in cases where the correct translation could not be found among the candidates, which consisted of checking the 30 strongest disambiguated features of an erroneously translated headword. We observed cases where the strongest features in the vectors are either very abstract and generic or too heterogeneous for our method to be able to perform well. This was the case with the headwords *characterisation*, *antecedent* and *thread*. In cases where the strongest features represented the concept clearly but the correct translation was not found, we examined cluster, WSD and MFT quality, as suggested by the parallel corpus. The main source of errors in these cases is the noise in the clusters which is often due to pre-processing errors, especially in the event of multi-word expressions. It seems that clustering is also problematic for abstract or generic words, where senses might be lumped together. The WSD step, on the other hand, does not seem to introduce noise to the procedure as it is correct in almost all the cases we have examined.

6 Discussion and conclusion

We have shown how cross-lingual WSD can be applied to bilingual lexicon extraction from comparable corpora. The disambiguation of source language features using translation clusters constitutes the main contribution of this work and presents several advantages. First, the method performs disambiguation by using sense descriptions derived from the data, which clearly differentiates our method from the approaches based on external lexicons and extends its applicability to resource-poor languages. The translation clusters acquired through WSI serve to disambiguate the features in the source language context vectors and to produce less noisy translated vectors. An additional advantage is that the sense clusters often contain more than one translation and, therefore, provide supplementary material for the comparison of the vectors in the target language.

The results show that data-driven semantic analysis can help to circumvent the need for an external seed dictionary, traditionally considered as a prerequisite for translation extraction from parallel corpora. Moreover, it is clear that disambiguating the vectors improves the quality of the extracted lexicons and manages to beat the simpler, but yet powerful, most frequent translation heuristic. These encouraging results pave the way towards pure data-driven methods for bilingual lexicon extraction. This knowledge-light approach can be applied to languages and domains that do not dispose of large-scale seed dictionaries but for which parallel corpora are available.

An avenue that we intend to explore in future work is to extract translations corresponding to different senses of the headwords. Up to now, research on translation extraction has most often aimed the identification of one good translation for a source word in the comparable corpus. This has also been the case because most works have focused on identifying translations for specialized terms that do not convey different senses. However, words in a general language corpus like Wikipedia can be polysemous and it is important to identify translations corresponding to their different senses. Moreover, polysemy makes the translation extraction procedure more difficult, as features corresponding to different senses are mingled in the same vector. A way to discover translations corresponding to different word senses would be to apply a monolingual WSI method on the source side of the comparable corpus which would group the closely related usages of the headwords together, and to then build vectors for each usage group hopefully describing a distinct sense. Using the generated sets of vectors separately will allow to extract translations corresponding to different senses of the source words.

References

Marianna Apidianaki, Nikola Ljubešić, and Darja Fišer. 2012. Disambiguating vectors for bilingual lexicon extraction from comparable corpora. In *Eighth Language Technologies Conference*, pages 10–15, Ljubljana, Slovenia.

Marianna Apidianaki. 2008. Translation-oriented sense induction based on parallel corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-08)*, pages 3269–3275, Marrakech, Morocco.

Marianna Apidianaki. 2009. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 77–85, Athens, Greece.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Hervé Déjean, Eric Gaussier, Jean-Michel Renders, and Fatiha Sadat. 2005. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2):111–124, February.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. In *TSD 2008 - Text Speech and Dialogue*, Lecture Notes in Computer Science, Brno, Czech Republic. Springer.

Pascale Fung. 1998. Machine translation and the information soup, third conference of the association for machine translation in the americas, amta '98, langhorne, pa, usa, october 28-31, 1998, proceedings. In *AMTA*, volume 1529 of *Lecture Notes in Computer Science*. Springer.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Amir Hazem and Emmanuel Morin. 2012. Ica for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC)*, Istanbul, Turkey.

Hiroyuki Kaji. 2003. Word sense acquisition from bilingual comparable corpora. In *HLT-NAACL*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *In Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Erwin Marsi and Emiel Krahmer. 2010. Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 752–760, Beijing, China, August. Coling 2010 Organizing Committee.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Pablo Gamallo Otero. 2007. Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of MT Summit XI*, pages 191–198.

Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Machine Translation Summit 2009*, page 8.

Reinhard Rapp. 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*, pages 519–526, College Park, Maryland, USA, June. Association for Computational Linguistics.

Xabier Saralegi, Iñaki San Vicente, and Antton Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the Building and using Comparable Corpora workshop, 6th International Conference on Language Resources and Evaluations (LREC)*, Marrakech, Morocco.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of Coling 2004*, pages 618–624, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Matthew G. Snover, Bonnie J. Dorr, and Richard M. Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *EMNLP*, pages 857–866.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Toma Erjavec, and Dan Tufi. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147.

Kun Yu and Junichi Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 121–124, Boulder, Colorado, June. Association for Computational Linguistics.