Vector Disambiguation for Translation Extraction from Comparable Corpora

Marianna Apidianaki
LIMSI-CNRS
Rue John von Neumann, F-91403, ORSAY CEDEX, France
E-mail: marianna@limsi.fr, http://www.limsi.fr/~marianna/

Nikola Ljubešić
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
E-mail: nikola.ljubesic@ffzg.hr, http://www.nljubesic.net/

Darja Fišer
Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
E-mail: darja.fiser@ff.uni-lj.si, http://lojze.lugos.si/darja

Keywords: word sense disambiguation, sense clustering, comparable corpora

Received: January 7, 2013

We present a new data-driven approach for enhancing the extraction of translation equivalents from comparable corpora which exploits bilingual lexico-semantic knowledge harvested from a parallel corpus. First, the bilingual lexicon obtained from word-aligning the parallel corpus replaces an external seed dictionary, making the approach knowledge-light and portable. Next, instead of using simple one-to-one mappings between the source and the target language, translation equivalents are clustered into sets of synonyms by a cross-lingual Word Sense Induction method. The obtained sense clusters enable us to expand the translation of vector features with several translation variants using a cross-lingual Word Sense Disambiguation method. Consequently, the vector features are disambiguated and translated with the translation variants included in the semantically most appropriate cluster, thus producing less noisy and richer vectors that allow for a more successful cross-lingual vector comparison than in previous methods.

Povzetek: V prispevku predstavljamo pristop za izboljšanje luščenja prevodnih ustreznic iz primerljivih korpusov z dodatnim virom leksiko-semantičnega znanja, izluščenega iz vzporednega korpusa.

1 Introduction

Due to the scarcity of general language parallel corpora, extracting translation information from comparable corpora has become a very active area of research in the past two decades. Identifying translation correspondences in comparable corpora offers low-resourced language pairs and domains a fast and affordable way to construct bilingual lexica and provides information useful for training Statistical Machine Translation systems (Munteanu and Marcu, 2005; Snover et al., 2008). The main idea behind translation extraction from comparable corpora is the assumption that a source word and its translation appear in similar contexts. n order to compare the context similarity of source and target words the same vector has to be produced, which means that the vectors of the one language have to be translated in the other language. Feature vector translation generally presupposes the availability of a bilingual dictionary (Fung, 1998; Rapp, 1999), which is however not the case for many language pairs or domains.

Another problem with the traditional approach to bilingual lexicon extraction and most of its extensions (Shao and Ng, 2004; Otero, 2007; Yu and Tsujii, 2009; Marsi and Krahmer, 2010) is that they neglect polysemy and consider a translation candidate as correct if it is an appropriate translation for at least one possible sense of the source word. This often corresponds to the most frequent sense of the word due to the way context vectors are built. An alternative to this consists in considering all translations provided for a source word in a bilingual dictionary but weighting them by their frequency in the target language (Prochasson et al., 2009; Hazem and Morin, 2012). The high quality of the information exploited by both these methods – generally found in hand-crafted resources – combined with the skewed distribution of the translations corresponding to different senses of the words, often leads to satisfying results. Nevertheless, this approach limits the usability of the proposed methods to languages and domains where such resources are available. We believe that relying on minimal resources that can be easily obtained for any language pair and domain, and combining them with automatic disambiguation of the features in the context vectors can lead to the production of cleaner vectors and, consequently, to higher quality results during lexicon extraction from comparable corpora.

The goal of this paper is twofold: first, we wish to eliminate the need for an external knowledge source by automatically extracting a bilingual lexicon from a parallel corpus. Second, we propose a way for disambiguating polysemous features in the context vectors, as these features may be translated differently according to the sense in which they are used in a given context.

The rest of the paper is organized as follows: In the next section, we present some related work on the subject. In Section 3, we present the resources that were used in our experiments. In Section 4, we describe the approach and the experimental setup in detail. The obtained results are presented and discussed in Session 5, after which the paper is wrapped up with some concluding remarks and ideas for future work.

2 Related work

The need to bypass pre-existing dictionaries has been addressed in several works on translation information extraction from comparable corpora. Koehn and Knight (2002) build the initial seed dictionary automatically, based on identical spelling features between the two languages (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 the seed lexicon approaches have been successfully combined by Fišer and Ljubešić (2011) who showed that the results with an automatically created seed lexicon that is 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 (EN) and Slovene (SL), which is the case in this experiment.

As for vector comparison, we believe we can produce less noisy vectors and improve their comparison across languages by using contextual information to disambiguate their features. This is done by a cross-lingual data-driven Word Sense Disambiguation method which assigns to each feature a cluster of semantically similar translations in the other language (Apidianaki, 2009). A similar idea has been implemented by Kaji (2003) who performed word clustering to extract sets of synonymous translation equivalents from from English-Japanese comparable corpora using pre-defined bilingual dictionaries. In addition, instead of providing one translation for each disambiguated feature, we translate it with all translation equivalents that belong to the assigned cluster similar to Déjean et al. (2005) who used a bilingual thesaurus instead of a lexicon.

The contribution of the work presented in this paper is a language independent and fully automated corpus-based approach to bilingual lexicon extraction from comparable corpora that does not rely on any external knowledge sources to determine word senses or translation equivalents.

3 Resources used

3.1 Comparable corpus

In this work, lexicon extraction is performed from a custom-built English-Slovene comparable corpus consisting of a collection of popular health and lifestyle articles from healthy-living magazines and the Internet. The core part of the corpus was collected manually from the Slovene reference corpus FidaPLUS (Arhar et al. 2007), already part-of-speech tagged and lemmatized. All the articles from the Slovene monthly health and lifestyle magazine (Zdravje) published between 2003 and 2005 have been included, amounting to one million words. For English, an equivalent amount of articles from the Health Magazine has been included. We PoS-tagged and lemmatized the English part of the corpus with TreeTagger (Schmid, 1994).

We then automatically extended the corpora from the two billion-word ukWaC (Ferraresi et al., 2008) and the 380 million-word slWaC (Ljubešić and Erjavec, 2011) that were constructed by crawling the .uk and .si domains. We took into account all the documents that pass a document similarity threshold with respect to the core corpus that was experimentally set in Fišer et al. (2011). The part of the extended corpus used in this experiment consists of 1 million words in each language.

3.2 Parallel corpus

3.2.1 Data

The information needed for applying our data-driven approach to the translation of source language vectors comes from an English-Slovene parallel corpus. Instead of an external seed lexicon used in most previous work, we translate source language vector features by exploiting the output of a cross-lingual WSD method (Apidianaki, 2009). The WSD method exploits the results of a cross-lingual Word Sense Induction (WSI) method that identifies word senses by clustering their translations in a parallel corpus. In the current setting, the English translations of Slovene words in a parallel corpus are clustered and the obtained sense clusters describe the senses of the source words.

The corpus used for sense induction is composed of the Slovene-English part of Europarl (release v6) (Koehn, 2005) and the Slovene-English part of the JRC-Acquis corpus (Steinberger et al., 2006), amounting to approximately 35M words per language.

So, the parallel corpus used for sense induction comes from a different domain than the comparable corpus described in Section 3.1. This is not the ideal scenario given that domain adaptation is important for the type of semantic processing we want to apply. There must be a noticeable shift in the senses present in the two corpora which makes the disambiguation stage harder and, in some cases, less interesting as true ambiguities become less frequent. The main reasons we opt for this configuration in this initial set of experiments are that there are very few large parallel corpora for the English-Slovene language pair, and that a comparable corpus and a gold standard needed for evaluation are available for the health domain. Furthermore, the combination of the two EU corpora provides sufficient material for training the unsupervised word sense induction and disambiguation methods that we intend to use. We should however note that, although the corpora pertain to different domains, they do contain a lot of general vocabulary. This is the case for both the EU corpus and the health domain corpus which is not medical (in the technical sense) but more popular, built from health and lifestyle magazines.

3.2.2 Pre-processing

Prior to being used for sense induction, the parallel corpus is subject to several pre-processing steps. We first eliminate sentence pairs with a great difference in length (i.e. cases where one sentence is more than three times longer than the corresponding sentence in the other language). Next, the corpus is lemmatized and PoS-tagged with the TreeTagger (for English) and the ToTaLe tool (for Slovene) (Erjavec et al., 2010). ToTaLe uses the TnT tagger (Brants, 2000) and was trained on MultextEast corpora (Erjavec, 2012). Two part-of-speech lexicons are built containing the PoS with which each word appears in the corpus. Next, the corpus is word-aligned with GIZA++ (Och and Ney, 2003) and two bilingual lexicons are extracted from the alignment results, one for each translation direction (EN–SL/SL–EN).

Several filters are then applied to clean the lexicons from noisy alignments. The translations are filtered on the basis of their alignment score (threshold: 0.01) and their PoS, keeping for each word only translations pertaining to the same grammatical category. We retain the intersecting alignments and use for clustering only translations that translate a source word more than 10 times in the training corpus. Even if this threshold leaves out some translations of the source words, it has the double merit of reducing data sparseness issues and eliminating erroneous translations which may be found in the lexicons because of spurious alignments. The filtered EN-SL lexicon contains entries for 6,384 nouns, 2,447 adjectives and 1,814 verbs with more than three translations in the training corpus. This lexicon is exploited for Word Sense Induction, as will be explained in Section 4.

3.2.3 Gold standard

We evaluate the results of the different experiments we carry out for extracting bilingual lexicons from comparable corpora by comparing them to a gold standard lexicon, which was comparable corpus and manually inspected. The gold standard lexicon contains 187 domain terms (nouns) that are present in the source language corpus with a minimum frequency of 50. Twenty-three of these terms have two attested translations in the corpus (e.g. EN *rectum* → SL *danka, rektum)* while the rest have just one (e.g. EN *breast* → SL *dojka).*

4 Experimental setup

4.1 Cross-lingual sense clustering

4.1.1 Vector building from the parallel corpus

The translations retained for each English target word (*w*) from the parallel corpus after the filtering process described in Section 3.2.2, are clustered on the basis of source language distributional information. Each Slovene translation *(T_i)* of *w* is characterized by a vector built from the co-occurrences of *w* in English. The vector contains the lemmas of content words (nouns, verbs and adjectives) that co-occur with *w* in the source side of the aligned sentences where it is translated by T_i, and their frequency counts. Using these vectors, pairwise similarities between the translations of *w* are calculated by a variation of the Weighted Jaccard measure (Grefenstette, 1994; Apidianaki, 2008).

For each translation T_i of *w*, let *N* be the number of features retained from the corresponding source context. Each feature F_j ($1 \leq j \leq N$) receives a total weight *$tw(F_j,T_i)$* with translation T_i defined as the product of the feature's global weight, *$gw(F_j)$*, and its local weight with that translation, *$lw(F_j,T_i)$*:

$$\mathrm{tw}(F_j, T_i) = \mathrm{gw}(F_j) \cdot \mathrm{lw}(F_j, T_i)$$

The global weight of a feature F_j depends on its dispersion in the contexts of *w*. More precisely, the global weight of the feature 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:

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

Each of the p_{ij}'s is computed as the ratio between the co-occurrence frequency of F_j with *w* when translated as T_i, denoted as *cooc_frequency(F_j,T_i)*, and the total number of features (*N*) seen with T_i:

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

Finally, the local weight *$lw(F_j,T_i)$* between F_j and T_i directly depends on their co-occurrence frequency:

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

4.1.2 Similarity calculation

The weights assigned to the features by the Weighted Jaccard measure reflect their relevance for calculating the similarity of the translation vectors. The score assigned to a pair of vectors indicates the degree of similarity of the corresponding translations. Translation pairs with a score above a threshold defined locally for each *w,* and dependent on the similarity scores assigned to its pairs of translations, are considered as semantically related.

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 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*)
	Adjectives	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)*

Table 1: Examples of nominal, verbal and adjectival entries from the English-Slovene sense cluster inventory.

The similarity threshold is set following the method proposed in Apidianaki and He (2010). This iterative procedure permits to define a local threshold for each *w* and to avoid using a static threshold that might not be appropriate for different words. The threshold (*T*) for a word *w* is initially set to the mean of the scores (above 0) of the translation pairs of *w*. The translation pairs of *w* are then divided into two sets (G_1 and G_2) according to whether they exceed, or are inferior to, the threshold. Then, the average of the scores of the translation pairs in each set is computed (m_1 and m_2) and a new threshold is created that is the average of m_1 and m_2 ($T = (m_1 + m_2)/2$). The new threshold serves to separate once again the translation pairs into two sets, a new threshold is calculated and the procedure is repeated until convergence is reached.

The similarity threshold calculated in this way permits to estimate the semantic proximity of the translations. Once this is done, the clustering algorithm groups the semantically similar Slovene translations into 'sense-clusters' describing the senses of the corresponding English words.

4.1.3 Translation clustering

The clustering algorithm takes as input the list of translations of the English word, their similarity scores and the similarity threshold, and outputs clusters of semantically related translations of the word in the target language. The clustering is performed in two steps. First, each translation pair with a similarity score exceeding the threshold is considered to have a pertinent relation and forms a cluster. The obtained two-element clusters might be enriched, during the second clustering step, by additional translations that are semantically related to all the translations already in the cluster. The clustering stops when all translations are included in some cluster and all their relations have been checked. All the elements in the final clusters are linked to each other by strong semantic relations, similar to cliques in undirected graphs.

Table 1 provides examples of clusters for English words of different PoS with clear sense distinctions in our training corpus. For each English word, we give the obtained clusters of Slovene translations, including a description of the sense described by each cluster.

For instance, the translations *krogla, sfera* and *področje* of the word *sphere* are grouped into two sense-clusters {*krogla*} and {*sfera, področje*} which describe the two senses of *sphere* observed in the corpus: “*geometrical shape*” and “*area*”. Similarly, the translations retained 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 (1518 nouns, 554 verbs and 513 adjectives) have more than one cluster.

4.2 Vector building from the comparable corpus

Context vectors in both the source and the target language are built for nouns occurring at least 50 times in the comparable corpus. This frequency threshold is required in order to obtain enough contextual data and ensure minimally reliable results in the lexicon extraction process.

As features in context vectors, we use three content words to the left and to the right of the retained nouns, stopping at the sentence boundary. The position of each content word is not taken into account, i.e. the context is seen as a bag of words. Our previous research (Fišer and Ljubešić, 2011; Ljubešić et al., 2011) has shown that encoding feature positions is mostly useful only when extracting translation candidates between closely related, syntactically similar languages.

Feature weights are calculated by the TF-IDF measure. TF is calculated as the relative frequency of a content word feature regarding all content word features in a specific context vector. IDF weights are calculated on the whole ukWaC and slWaC corpora in a typical IR manner by obeying document boundaries. Our previous research (Ljubešić et al., 2011) has shown that TF-IDF feature weights perform as good as the more complex log-likelihood weighting and better than pure relative frequency. These feature weights serve additionally to filter out ‘weak’ features that are shown not to be useful for the lexicon extraction task (see Section 5.2).

4.3 Vector disambiguation

4.3.1 A data-driven approach

In order to identify the translations of the source words in the target language side of the comparable corpus, the vectors built in the two languages must be compared. This comparison serves to quantify the similarity of the source and target language words represented by the vectors, and the highest ranked pairs are proposed as entries for the lexicon.

As the vectors have been built from monolingual corpora, the source language vectors must first be translated into the target language. As explained above, in most previous work on bilingual lexicon building from comparable corpora, the vectors were translated using external seed dictionaries. The first translation proposed for a word in the dictionary was used to translate all the instances of the word in the vectors irrespective of their sense, and no disambiguation was performed.

The use of external resources ensures the quality of the translations used for translating the source vectors. Moreover, the selection of the most frequent translation often results in good translations because of the skewed distribution of the translations corresponding to different senses of the words. Nevertheless, this technique limits the usability of the proposed lexicon extraction methods to languages and domains where such resources are available.

In this work, instead of using an external bilingual dictionary, we translate the source language vectors using the data-driven cross-lingual WSD method proposed by Apidianaki (2009). The method exploits the sense clusters acquired from parallel corpora by the sense induction method described in Section 4.1. This property extends the applicability of the method to languages lacking large-scale lexical resources but for which parallel corpora are available.

4.3.2 Cross-lingual WSD

The sense clusters of translations obtained during sense induction (cf. Section 4.1) represent the candidate senses of the English words in the parallel corpus. We exploit this sense inventory for disambiguating the features in the English vectors that were extracted from the comparable corpus. More precisely, the WSD method has to select for each feature in the vectors built from the comparable corpus, the cluster that correctly translates its sense in the target language.

In the current setting, the selection is performed by comparing information from the context of the vector features to the distributional information that served to estimate the semantic similarity of the clustered translations. The context of a feature to be disambiguated corresponds to the rest of the vector where it appears. Inside the vectors, the features are ordered according to their weight (calculated as explained in Section 4.1). The feature weights serve to filter out the *weak* features (i.e. features with a score below a threshold) which were shown not to be useful for the lexicon extraction task. The threshold was experimentally set at 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 built from the parallel corpus which revealed the semantic similarity of the clustered translations. If common features (*CF'*s) are found between the context of a feature and just one cluster, this cluster is selected to describe the feature's sense. Otherwise, if there exist *CF'*s with more than one cluster, then a score is assigned to each ‘cluster-feature’ association. This weight corresponds to the mean of the weights of the *CF's* relative to the clustered translations (weights assigned to each feature during clustering). In the following formula, CF_j is the set of *CF'*s found between the cluster and the new context and N_{CF} is the number of translations T_i in the cluster characterized by a *CF*:

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

The highest scored cluster is selected and assigned to the feature as a sense tag. The features are also tagged with the most frequent (MF) translation of the word in the parallel training corpus, which sometimes already exists in the cluster selected during WSD.

In Table 2, we present some examples of disambiguated vector features of different PoS. For each case, we provide: the headword entry to which the vector corresponds; a feature from the vector that has been disambiguated (a noun, a verb and an adjective, respectively, in the three examples); and the context that was used for disambiguation, which consists of the other strong features found in the same vector (i.e. features with a weight above the threshold). From the candidate clusters available for the feature (given in column 4), the WSD method selects the most appropriate one (in boldface) to describe the feature's sense in this context. In the last column of the table, we provide the most frequent sense/translation (MF) for the feature.

Headword	Feature (PoS)	Context	Candidate clusters	MF alignment
infertility	treatment (n)	*doctor, diabetes, health, emergency, check, ...*	- {**zdravljenje, obdelava, obravnavanje, obravnava, ravnanje}** (*treat an illness)* - {čiščenje} (*treat a person/animal)* - {raba} (*usage*)	obravnava
clot	seal (v)	*block, heart, vessel, pressure, infection, ...*	- {**tesniti**} (*to be waterproof or airtight)* - {zapreti, zapečatiti} (*to close)*	zapečatiti
arrhythmia	irregular (a)	*heart, abnormal, monitor, failure, risk, ...*	- {**nepravilen, nereden}** (*not regular)* - {ilegalen} (*illegal*)	nepravilen

Table 2: Disambiguation results.

We observe that the MF translation may already exist in the cluster selected by the WSD method, like in the first example where *obravnava* is already in the selected cluster. The inverse, i.e. that the MF is not found in the proposed cluster, is also possible as is the case with the *zapečatiti* translation of the verb *seal.*

The disambiguation of source language features using cross-lingual sense 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 extends its applicability to resource-poor languages. This procedure clearly differentiates our method from previous approaches where the first translation in a dictionary – which is often the most frequent one – was selected for translating each vector feature. An additional advantage is that the sense clusters assigned to features may contain more than one translation. This property is important in this setting as it provides supplementary material for the comparison of the vectors in the target language.

Cross-lingual vector comparison

The translation of the source vectors into the target language, performed as described in the previous section, makes possible the comparison of the vectors in the same vector space. We experiment with three different ways of translating features:

1. by keeping the translation a feature was most frequently aligned to in the parallel corpus (MF);
2. by keeping the most frequent translation from the cluster assigned to the feature during disambiguation (CLMF); 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 is used as a baseline since instead of the sense clustering and WSD results, it just uses the "most frequent sense/alignment" heuristic. In the first batch of the experiments, we noticed that the results of the CL approach heavily depend on the part-of-speech of the features. So, we divided the CL approach into three sub-approaches:

1. translate only nouns with the clusters and other features with the MF approach (CL-n);
2. translate nouns and adjectives with the clusters and verbs with the MF approach (CL-na); and
3. translate all PoS with the clusters (CL-nav).

The distance between the translated source and the target-language vectors is computed by the Dice metric which has proven to be very efficient when combined with the TF-IDF weighting (Ljubešić et al., 2011).

During our experiments, we noticed that discarding the weakest features from the context vectors in the source language significantly improves the results. So, we also experiment with a minimum feature weight threshold and call this parameter the 'minimum feature weight threshold' (mfwt). 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 and discussion of the results

5.1 Evaluation setting

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

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

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.

Since most of the entries in our gold standard contain just one translation, we did not consider using more advanced evaluation measures for ranked results, like mean average precision (MAP).

5.2 Results and discussion

The results of our final experiment are shown in Figure 1. The x axis shows the minimum feature weight threshold (mfwt) while on the y axis the evaluation measure MRR is plotted.

The phenomenon that is first observed in the graph is the one for which we have introduced the minimum feature weight threshold parameter: the best results are obtained when discarding all features that have a TF-IDF weight score lower than 0.01. This is something we had not noticed before and that we intend to explore more thoroughly in a new set of experiments, by measuring its consistency when different weight measures, distance measures, seed lexicons, language pairs and comparable corpora are used.

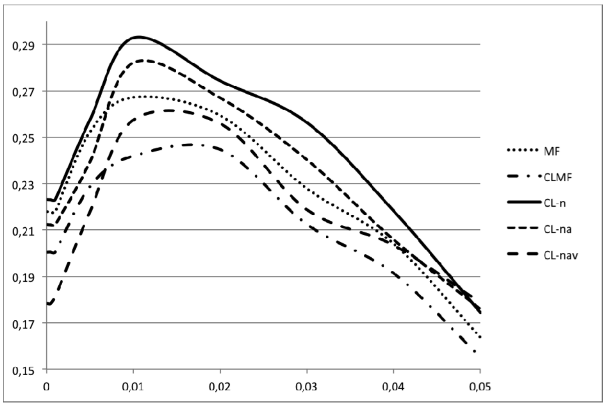

Figure 1: Evaluation of different approaches to lexicon extraction.

The lowest results are consistently obtained when using the CLMF approach, which consists in using only the most frequent translation from the cluster chosen through the WSD procedure. A possible reason for this is the fact that alignment frequencies used for finding the most frequent translation in the cluster were calculated on a corpus of a different domain than our comparable corpus (Europarl vs. health corpus).

The baseline which always uses the most frequent translation of the feature from the parallel corpus, without sense clustering and WSD, achieves a medium result. The baseline is outperformed by the CL-n and the CL-na approaches but performs better than the CL-nav approach, which shows that taking verbs into account deteriorates the quality of the results.

The different CL approaches yield somewhat expected results. The biggest gain is obtained from clustering and WSD information calculated on nouns, nouns and adjectives scored second and the lowest results are obtained when verbs are added to the mix. This is probably due to the fact that the verbal clusters are noisier than the nominal and adjectival ones. We intend to further explore this issue.

Since our gold standard is quite small, we checked the statistical significance of the difference in the results of the baseline MF approach and the winning CL-n approach. We used the approximate randomization procedure with R = 1000 (i.e. 1000 random assignments were done without replacement of the two sets of results). The resulting *p-value* is 0.091, which is higher than the commonly used 0.05 threshold.

These results show that in our future experiments we will need a larger gold standard to draw safer conclusions on the statistical significance of the results. However, since the p-value is below 0.1 and is accompanied by a consistent increase in performance throughout a large number of experiments, we are rather confident that this increase is not the result of random variation.

The main conclusions that can be drawn from the reported results here are the following:

- extending the feature set with multiple translations obtained by sense clustering and word sense disambiguation of features is beneficial to the lexicon extraction procedure;
- the most valuable information obtained from the clustering and WSD approach comes from nouns;
- using just the most frequent translation inside the cluster selected during WSD does not yield good results; and
- further investigation of the improvement that occurs when weak features are discarded is needed.

6 Conclusions and future work

We presented an approach that allows the use of lexico-semantic knowledge acquired from parallel corpora to improve the extraction of translation equivalents from comparable corpora. A parallel corpus served as the source of the seed dictionary, so that no external knowledge source is needed for the translation of features in context vectors. In addition, the seed dictionary was enhanced with clusters of translation variants obtained from the parallel corpus in an unsupervised way. The cross-lingual clusters were used to disambiguate the features in the context vectors, reducing noise, and allowed for a more accurate comparison of source and target vectors. Furthermore, the tagging of the vector features with clusters during disambiguation increased the translation information available for each feature and, therefore, facilitated the comparison of context vectors across languages.

The results show that lexico-semantic knowledge derived from a parallel corpus can help to circumvent the need for an external seed dictionary, traditionally considered as a pre-requisite for bilingual lexicon extraction from parallel corpora. Moreover, disambiguating the vectors improves the quality of the extracted lexicons and manages to beat the simpler, if powerful, most frequent sense/alignment heuristic.

These encouraging results pave the way towards pure data-driven methods for bilingual lexicon extraction from comparable corpora. 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. Moreover, the use of a data-driven cross-lingual WSD method, such as the one proposed in this paper, can contribute to obtain less noisy translated vectors, which is important especially when lexicon extraction is performed from general language comparable corpora.

The experiments carried out till now focus on a health comparable corpus. Although this is not a very specialized corpus but a rather popular one, cases of true polysemy are still less frequent than in a general corpus.

We would thus like to extend this work by applying the method to a more general comparable corpus, for instance a corpus built from Wikipedia texts. We expect that the effect of applying the WSD method on a general corpus will be highly beneficial, as ambiguity problems will be more prevalent.

We also want to explore the use of second order co-occurrences for disambiguation. For the moment, the context used to disambiguate vector features consists of other features that appear in the same vector. However, these features are direct co-occurrences of the headword, which does not necessarily mean that the features themselves co-occur with each other in the corpus. We consider that it would be preferable to replace this context with the co-occurrences of the features in the corpus for disambiguation, which would correspond to the second order co-occurrences of the English words, and investigate the effect of using this type of context on lexicon extraction.

References

[1] Marianna Apidianaki (2008) Translation-oriented Word Sense Induction based on Parallel Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

[2] Marianna Apidianaki and Yifan He (2010) An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT),* Paris, France, 219–226.

[3] 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)*, Athens, Greece, 77–85.

[4] Špela Arhar, Vojko Gorjanc and Simon Krek (2007) FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools. In *Proceedings of the Corpus Linguistics conference*, Birmingham, UK.

[5] Thorsten Brants (2000) TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP),* Seattle, WA.

[6] 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.

[7] Tomaž Erjavec (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

[8] Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. (2010) The JOS linguistically tagged corpus of Slovene. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC)*,Valletta, Malta.

[9] Adriano Ferraresi, Eros Zanchetta, Marco Baroni and Sylvia Bernardini (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4): Can we beat Google?,* Marrakech, Morocco, 47–54.

[10] Darja Fišer and Nikola Ljubešić (2011) Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP),* Hissar, Bulgaria, 125–131.

[11] Darja Fišer, Nikola Ljubešić, Špela Vintar and Senja Pollak (2011) Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th BUCC Workshop: Comparable Corpora and the Web,* Portland, Oregon, USA, 19–26.

[12] Pascale Fung (1998) Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora. *Lecture Notes in Artificial Intelligence*, Springer, Vol. 1529, 1–17.

[13] Gregory Grefenstette (1994) *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Norwell, MA.

[14] Amir Hazem and Emmanuel Morin (2012) ICA for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 5th Building and Using Comparable Corpora (BUCC) workshop,* Istanbul, Turkey, 126-133.

[15] Hiroyuki Kaji (2003) Word sense acquisition from bilingual comparable corpora. In *Proceedings of HLT-NAACL*, Edmonton, Canada, 32–39.

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

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

[18] Nikola Ljubešić and Tomaž Erjavec (2011) hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Proceedings of Text, Speech and Dialogue (TSD),* Lecture Notes in Computer Science (LNCS) Vol. 6836, Springer, 395–402.

[19] Nikola Ljubešić, Darja Fišer, Špela Vintar and Senja Pollak (2011) Bilingual lexicon extraction from comparable corpora: A comparative study. In *Proceedings of the International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia.

[20] 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),* Beijing, China, 752–760.

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

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

[23] Pablo Gamallo Otero (2007) Learning bilingual lexicons from comparable English and Spanish corpora. In *Proceedings of Machine Translation (MT) Summit XI*, Copenhagen, Denmark, 191–198.

[24] Emmanuel Prochasson, Emmanuel Morin and Kyo Kageura (2009) Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Machine Translation Summit (MT Summit XII)*, Ottawa, Ontario, Canada, 284-291.

[25] 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 (ACL),* College Park, Maryland, USA, 519–526.

[26] Xabier Saralegi, Iñaki San Vicente, Antton Gurrutxaga (2008) Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the 1st Building and Using Comparable Corpora (BUCC) workshop,* Marrakech, Morocco.

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

[28] Li Shao and Hwee Tou Ng (2004) Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING),* Geneva, Switzerland, 618–624.

[29] Matthew Snover, Bonnie Dorr and Richard Schwartz (2008) Language and Translation Model Adaptation using Comparable Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawai, 857-866.

[30] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş and Dániel Varga (2006) The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2142–2147.

[31] Kun Yu and Junichi Tsujii (2009) Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. In *Proceedings of NAACL/HLT 2009*, Boulder, Colorado, USA, 121–124.