Disambiguating vectors for bilingual lexicon extraction from comparable corpora

Marianna Apidianaki,* Nikola Ljubešić,† Darja Fišer[#]

* LIMSI-CNRS
BP 133, F-91403 ORSAY CEDEX, France
marianna.apidianaki@limsi.fr
† Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr
[#] Department of Translation, Facult of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Abstract

This paper presents an approach to enhance the extraction of translation equivalents from comparable corpora by plugging in 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 1:1 mappings between the source and the target language, translation equivalents are clustered into sets of synonyms based on contextual similarities, enabling us to expand the translation of vector features with several translation variants. And last but not least, the vector features are disambiguated and translated only with the translation variants from the most appropriate cluster, thus producing less noisy vectors that allow for a more successful cross-lingual comparison of the vectors compared to simpler methods.

Razdvoumljanje vektorjev za izboljšanje luščenja dvojezičnih leksikonov iz primerljivih korpusov

V prispevku predstavljamo pristop za izboljšanje luščenja prevodnih ustreznic iz primerljivih korpusov z dodatnim virom leksiko-semantičnega znanja, izluščenega iz vzporednega korpusa. Za razliko od večine sorodnih pristopov dvojezični leksikon, potreben za prevajanje kontekstnih vektorjev, izdelamo avtomatsko iz vzporednega korpusa. Tako pristop ni več odvisen od slovarja, potrebnega za prevajanje kontekstnih vektorjev in je tako prenosljiv na številne jezikovne pare in strokovna področja. V naslednjem koraku prevodne ustreznice v dvojezičnem leksikonu razvrstimo v gruče, kar nam omogoča, da lastnosti v kontekstnih vektorjev, izdelanih iz primerljivih korpusov, prevajamo z več kot eno prevodno različico. To nam olajša primerjavo kontekstnih vektorjev v izvornem in ciljnem jeziku. Tretja izboljšava, ki jo v prispevku predstavljamo, pa je razdvoumljanje večpomenski lastnosti kontekstnih vektorjev iz primerljivega korpusa z gručami, generiranimi iz dvojezičnega leksikona, ki omogoča natančnejše prevajanje vektorjev in izboljša njihovo primerjavo z vektorji v ciljnem jeziku.

1. Motivation and related work

Due to the lack of general language parallel corpora, finding translations in comparable corpora has become a very active area of research. The main idea behind this approach is the assumption that a source word and its translation appear in similar contexts, so that in order to identify them their contexts are compared via a seed dictionary (Fung, 1998; Rapp, 1999). The biggest advantage of the approach is that it offers low-resourced language pairs and domains a fast and affordable way to construct bilingual lexica. However, it also presupposes the availability of a bilingual dictionary to translate vector features, which is not the case for many language pairs or domains. In addition, the original approach and most of its extensions (Shao and Ng, 2004; Otero, 2007; Yu and Tsujii, 2009; Marsi and Krahmer, 2010) 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, which will often be the most frequent sense of the word due to the way context vectors are built.

The goal of this paper is twofold: (1) we eliminate the need for an external knowledge source by automatically extracting a bilingual lexicon from a parallel corpus, and (2) we propose a way of 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 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. Cognate detection has also been used by Saralegi et al. (2008). Both approaches have been successfully combined by Fišer and Ljubešić (2011) who showed that the results with an automatically created seed lexicon, based on the similarity between the languages, can be as good as with a pre-existing dictionary.

But all these approaches 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. We believe we can produce less noisy vectors and improve their comparison across languages by using contextual information to disambiguate their features. A similar idea has been implemented by Kaji (2003) who clustered synonymous Japanese translations of English words in 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 use a bilingual thesaurus instead of a lexicon. The contribution of 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.

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

2. Resources used

2.1. Comparable corpus

The custom-built English-Slovene comparable corpus that we use for bilingual lexicon extraction is a collection of popular health and lifestyle articles found in healthy-living magazines and on 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 (PoS) tagged and lemmatized. All articles from the Zdravje magazine published between 2003 and 2005 have been included, amounting to 1 million words. For English, an equivalent amount of articles from Health Magazine has been included. We PoS-tagged and lemmatized the English part of the corpus with TreeTagger (Schmid, 1994).

We then extended the initial corpus automatically from the 2 billion-word ukWaC (Ferraresi et al., 2008) and the 380 million-word slWaC (Ljubešić and Erjavec, 2011). 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).

2.2. Parallel corpus

2.2.1. Data

In this work, we enhance bilingual lexicon extraction from comparable corpora by applying a data-driven approach to the translation of source vectors. More precisely, we replace the external seed lexicon, used in previous work on lexicon extraction from comparable corpora, with the output of a cross-lingual WSD method (Apidianaki, 2009). The method exploits the results of a cross-lingual Word Sense Induction (CL-WSI) method that identifies word senses by clustering their translations in a parallel corpus. In the current setting, the English translations of Slovene in a parallel corpus are clustered and the obtained sense-clusters describe the senses of 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.

2.2.2. Pre-processing

Prior to being used for sense induction, the training corpus is subject to several pre-processing steps, such as elimination of sentence pairs with a great difference in length, lemmatization and PoS-tagging with 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 from the alignment results, one for each translation direction (EN–SL/SL–EN).

The lexicons are cleaned by applying a set of filters, in order to retain only intersecting alignments of the same PoS. 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 translations used for clustering are the ones with a minimum frequency of 10 in the training corpus and a minimum alignment certainty of 0.01. The resulting lexicon is then used for word sense induction (cf. Section 3).

2.3. Gold standard

We evaluate the results of different experimental settings by comparing them to a gold standard lexicon, which was collected from the corpus and manually inspected. It contains 187 domain terms (nouns) that are present in the source language corpus with a minimum frequency of 50. 23 of these terms have two attested translations in the corpus (e.g. Eng. *rectum* → Slo. *danka, rektum)* while the rest have just one (e.g. Eng. *breast* → Slo. *dojka).*

3. Experimental setup

3.1. Cross-lingual sense clustering

The translations of each English content word (*w*) in the parallel corpus are clustered on the basis of source language distributional information. Each Slovene translation (*t*) of *w* is characterized by a vector built from the co-occurrences of *w* in the aligned sentences where it is translated by *t.* The vectors contain lemmas of content words that co-occur with *w* 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). This measure assigns weights to the features that reflect their relevance for calculating the similarity of the vectors. The score assigned to a pair of vectors and the corresponding translations indicates their degree of similarity. 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.[1]

The clustering algorithm groups Slovene translations into clusters that describe the senses of the corresponding English words. More precisely, the algorithm takes as input the list of translations of an English word, their similarity scores and the computed similarity threshold, and it outputs clusters that contain semantically related translations. Table 1 gives examples of clusters for words of different PoS with clear sense distinctions. For each English word, we provide its clusters of Slovene translations that were obtained and include a description of the sense described by each cluster. For instance, the clusters of the word *sphere*: *{krogla}* and *{sfera, področje}*, describe the two senses of *sphere* observed in the corpus: "*geometrical shape*" and "*area*". The obtained cluster inventory contains 13,352 clusters for 8,892 words. 2,585 of the words (1518 nouns, 554 verbs and 513 adjectives) have more then one cluster.

[1] The threshold is set following the method proposed in Apidianaki and He (2010).

<table>
<tr><th>Language</th><th>POS</th><th>Source word</th><th>Slovene sense clusters</th></tr>
<tr><td rowspan="12">EN-SL</td><td rowspan="4">Nouns</td><td>sphere</td><td>{krogla} (geometrical shape)
{sfera, področje} (area)</td></tr>
<tr><td>address</td><td>{obravnava, reševanje, obravnavanje} (dealing with)
{naslov} (postal address)</td></tr>
<tr><td>portion</td><td>{kos} (piece)
{obrok, porcija} (serving)
{delež} (share)</td></tr>
<tr><td>figure</td><td>{številka, podatek, znesek} (amount)
{slika} (image)
{osebnost} (person)</td></tr>
<tr><td rowspan="4">Verbs</td><td>seal</td><td>{tesniti} (to be water-/airtight)
{zapreti, zapečatiti} (to close an envelope or other container)</td></tr>
<tr><td>weigh</td><td>{pretehtati} (consider possibilities)
{tehtati, stehtati} (check weight)</td></tr>
<tr><td>educate</td><td>{poučiti} (give information)
{izobraževati, izobraziti} (give education)</td></tr>
<tr><td>consume</td><td>{potrošiti} (spend money/goods)
{uživati, zaužiti} (eat/drink)</td></tr>
<tr><td rowspan="4">Adjs</td><td>mature</td><td>{zrel, odrasel} (adult)
{zorjen, zrel} (ripe)</td></tr>
<tr><td>minor</td><td>{nepomemben} (not very important)
{mladoleten, majhen} (under 18 years old)</td></tr>
<tr><td>juvenile</td><td>{nedorasel} (not adult/biologically mature yet)
{mladoleten, mladoletniški} (not 18/legally adult yet)</td></tr>
<tr><td>remote</td><td>{odmaknjen, odročen} (far away and not easily accessible)
{oddaljen, daljinski} (controlled from a distance)</td></tr>
</table>

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

3.2. 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. As features, 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. Feature weights are calculated by the TF-IDF measure, where IDF weights are calculated on the whole ukWaC and slWaC. The feature weights serve to filter out the *weak* features that were shown not to be useful for the lexicon extraction task. The threshold is experimentally set at 0.01.

3.3. Vector disambiguation

3.3.1. A data-driven approach

For extracting bilingual lexicons from comparable corpora, 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. For that, source language vectors must first be translated into the target language. In most previous work, the vectors were translated with external dictionaries: the first translation 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, we translate the source language vectors using a data-driven cross-lingual WSD method (CL-WSD) (Apidianaki, 2009). The method exploits the sense clusters acquired from parallel corpora (see Section 3.1). This property extends the applicability of the method to languages lacking large-scale lexical resources but for which parallel corpora are available.

3.3.2. Cross-lingual WSD

The sense clusters of translations obtained during WSI 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 extracted from the comparable corpus. More precisely, the CL-WSD method has to select among the available clusters the one that correctly translates in Slovene the sense of the English features contained in the vectors built from the comparable corpus.

In the current setting, the selection is performed by comparing information from the context of the 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 ranked and filtered according to their score (calculated as explained in Section 3.2). The retained features are 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 semantic similarity of the clustered translations.

If common features (CFs) 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 CFs 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 *CFs* relative to the clustered translations (weights assigned to each feature during WSI). In the following formula, CF_j is the set of *CFs* 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 translation of the word in the 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 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 a threshold). From the candidate clusters available for the feature (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. 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.

3.4. Cross-lingual vector comparison

For context vectors to be comparable between languages, the same vector space has to be produced. This is done by translating the source language features to the target language. We translated the features in three ways:

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. Since 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, 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).

Once the source language vectors are built, 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). We also experiment with a minimum feature weight threshold since, during our experiments, we observed the phenomenon where discarding the weakest features from the context vectors in the source language significantly improves the results. We 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.

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

Table 2: Examples of disambiguated vector features.

4. Evaluation and discussion of the results

4.1. Evaluation procedure

We evaluate the final result of our method, i.e. the ranked lists of translation candidates for gold standard entries by the mean reciprocal rank (MRR) 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 of our gold standard contain just one translation, we did not consider using some more advanced evaluation measure, like mean average precision.

4.2. Results & 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 on the graphs 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 we will look into this phenomenon 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.

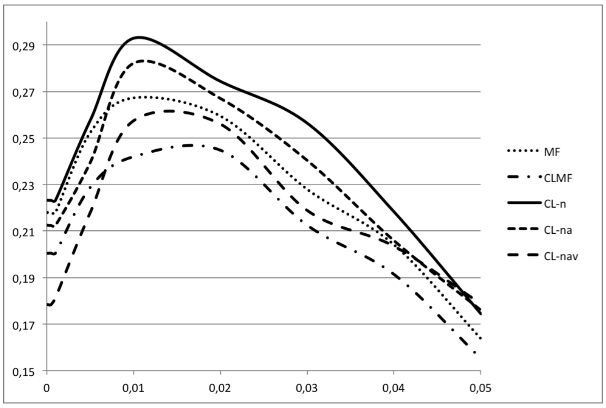

Figure 1: Evaluation of the different translation approaches regarding the minimum feature weight threshold.

Overall, the worst results are obtained when using the CLMF approach, i.e. 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 alignment in the cluster were calculated on a corpus of a different domain than our comparable corpus (Europarl vs. health corpus). The baseline that always uses the most frequent translation of the feature from the parallel corpus, without any sense clustering and WSD, achieves a medium result, being outperformed by the CL-n and the CL-na approach but outperforming the CL-nav approach.

The CL sub-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 explore this issue in future work.

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 results demonstrated here are that:

- 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 phenomenon where discarding the weak features improves the result is needed.

5. Conclusions and future work

We presented an approach that allows to use lexico-semantic knowledge acquired from parallel corpora in order to improve the extraction of translation equivalents from comparable corpora. A parallel corpus served as the source of the seed dictionary, so that the translation of features in context vectors no longer relies on an external knowledge source. 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, thus 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 prerequisite for bilingual lexicon 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 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.

Another avenue that we want to explore is to use second order co-occurrences for disambiguation. For the moment, the context used to disambiguate vector features consists of the 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 nouns, and investigate the effect of using this type of context on lexicon extraction. Last but not least, we would like to apply the method to the opposite direction (i.e. from Slovene to English) and compare the results obtained in both directions.

6. References

M. Apidianaki. 2008. Translation-oriented Word Sense Induction based on Parallel Corpora. *Proc. of LREC*, Marrakech, Morocco.

M. Apidianaki, Y. He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. *Proc. of the 7th IWSLT,* 219–226, Paris, France.

M. Apidianaki. 2009. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. *Proc. of the 12th EACL-09,* 77–85, Athens, Greece.

Š. Arhar, V. Gorjanc, S. Krek 2007. FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools, *Proc. of the Corpus Linguistics conference*, Birmingham, UK.

T. Brants. 2000. Tnt a statistical part-of-speech tagger. In *Proceedings of the 6th ANLP,* Seattle, WA.

H. Déjean, E. Gaussier, J. Renders, F. Sadat. 2005. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *Artif. Intell. Med.,* 33(2):111–124.

T. Erjavec, D. Fišer, S. Krek, N. Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. *Proc. of 7th LREC,* Valletta, Malta.

A. Ferraresi, E. Zanchetta, M. Baroni, S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. *Proc. of the 4th WAC: Can we beat Google,* 47–54.

D. Fišer, N. Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. *Proc. of RANLP,* 125–131, Hissar, Bulgaria.

D. Fišer, N. Ljubešić, Š. Vintar, S. Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. *Proc. of the 4th BUCC: Comparable Corpora and the Web,* 19–26, Portland, Oregon, USA.

P. Fung. 1998. Machine translation and the information soup, third conference of the association for machine translation in the Americas, *AMTA, Vol. 1529 of LNCS.* Springer.

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

H. Kaji. 2003. Word sense acquisition from bilingual comparable corpora. *Proc. of HLT-NAACL.*

P. Koehn, K. Knight. 2002. Learning a translation lexicon from monolingual corpora. *Proc. of ACL Workshop on Unsupervised Lexical Acquisition,* 9–16.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proc. of MT Summit X,* 79–86, Phuket, Thailand.

N. Ljubešić, T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. *TSD Vol. 6836 of LNCS,* 395–402. Springer.

N. Ljubešić, D. Fišer, Š. Vintar, S. Pollak. 2011. Bilingual lexicon extraction from comparable corpora: A comparative study. *Proc. of WOLER*, Ljubljana, Slovenia August 1-5, 2011.

E. Marsi, E. Krahmer. 2010. Automatic analysis of semantic similarity in comparable text through syntactic tree matching. *Proc. of COLING,* 752–760, Beijing, China, August.

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

P. Gamallo Otero. 2007. Learning bilingual lexicons from comparable English and Spanish corpora. *Proc. of MT Summit XI,* 191–198.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. *Proc. of the 37th ACL,* 519–526, College Park, Maryland, USA.

Z. Saralegi, I. San Vicente, A. Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proc. of the 1st Building and Using Comparable Corpora (BUCC) workshop,* Marrakech, Morocco.

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

L. Shao, H. Tou Ng. 2004. Mining new word translations from comparable corpora. *Proc. of COLING,* 618–624, Geneva, Switzerland, Aug 23–Aug 27.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proc. of the 5th LREC,* 2142–2147.

K. Yu, J. Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. *Proc. of ACL,* 121–124, Boulder, Colorado, USA.