Bilingual Lexicon Extraction from Comparable Corpora for Closely Related Languages

Darja Fišer
Faculty of Arts,
Univeristy of Ljubljana
darja.fiser@ff.uni-lj.si

Nikola Ljubešić
Faculty of Humanities and Social Sciences,
University of Zagreb
nikola.ljubesic@ffzg.hr

Abstract

In this paper we present a knowledge-light approach to extract a bilingual lexicon for closely related languages from comparable corpora. While in most related work an existing dictionary is used to translate context vectors, we take advantage of the similarities between languages instead and build a seed lexicon from words that are identical in both languages and then further extend it with context-based cognates and translations of the most frequent words. We also use cognates for reranking translation candidates obtained via context similarity and extract translation equivalents for all content words, not just nouns as in most related work. The results are very encouraging, suggesting that other similar languages could benefit from the same approach. By enlarging the seed lexicon with cognates and translations of the most frequent words and by cognate-based reranking of translation candidates we were able to improve the average baseline precision from 0.592 to 0.797 on the mean reciprocal rank for the ten top-ranking translation candidates for nouns, verbs and adjectives with a 46% recall on the gold standard of 1000 random entries from a traditional dictionary.

1 Introduction

Most cross-lingual NLP applications require bilingual lexicons but their compilation is still a major bottleneck in computational linguistics. Automatic extraction of bilingual lexicons is typically performed on parallel corpora (Och and Ney, 2000) but they exist only for a limited number of language pairs and domains and it is often impractical or even impossible to build one from scratch.

This is why an alternative approach has become popular in recent years. It relies on texts in two languages which are not parallel but comparable (Fung, 1998; Rapp, 1999) and therefore easier to compile, especially from the increasingly rich web data (Xiao and McEnery, 2006). The approach relies on the assumption that the term and its translation appear in similar contexts (Fung, 1998; Rapp, 1999). This means that the translation of a source word can be found by identifying a target word which has the most similar context vector in a comparable corpus. However, a direct comparison of vectors in two different languages is not possible, which is why a dictionary is needed to first translate the features of source context vectors into the target language and compute similarity on those. But this step seems paradoxical: the very reason why we are applying the complex comparable corpus technique for extracting translation equivalents is the fact that we do not have a bilingual dictionary at our disposal in the first place. This issue has largely remained unaddressed in previous research, which is why we propose a knowledge-light approach that does not require any bilingual resource. Instead, it takes advantage of similarities between the source and the target language in order to obtain a seed lexicon used for translating features of context vectors.

The paper is structured as follows: in the following section we give an overview of related work. In Section 3 we present the construction of the resources used in the experiment. Section 4 describes the experimental setup and reports the results of automatic and manual evaluation. We conclude the paper with final remarks and ideas for future work.

2 Related Work

The seminal papers on bilingual lexicon extraction from non-parallel texts are (Fung, 1998) and (Rapp, 1999) whose main assumption is that the

Proceedings of Recent Advances in Natural Language Processing, pages 125–131,
Hissar, Bulgaria, 12-14 September 2011.

term and its translation share similar contexts. The method consists of two steps: first, contexts of words are modeled and then similarity between the source-language and target-language contexts are measured with the help of a dictionary. Most approaches represent contexts as weighted collections of words using log-likelihood (Ismail and Manandhar, 2010), TF-IDF (Fung, 1998) or PMI (Shezaf and Rappoport, 2010). After building context vectors for words in both languages, the similarity between a source word's context vector and all the context vectors in the target language is computed using a similarity measure, such as cosine (Fung, 1998), Jaccard (Otero and Campos, 2005) or Dice (Otero, 2007).

If we want to compare context vectors across languages, the translation of features in context vectors is required, which assumes that a dictionary is available. Alternative solutions for situations when this is not the case have not been explored to a great extent but (Koehn and Knight, 2002) show that it is possible to obtain a seed lexicon from identical and similarly spelled words that is directly extracted from the comparable corpus. Taking the idea one step further, (Al-Onaizan and Knight, 2002) and (Shao and Ng, 2004) use transliteration rules for Arabic and Chinese respectively to harvest translation candidates, which is especially efficient for named entities and new vocabulary not yet present in dictionaries. At the subword level, (Markó et al., 2005) defined a set of string substitution rules to obtain domain-specific Spanish-Portugese cognates. As an addition to the standard approach, (Saralegi et al., 2008) use string similarity as a reranking criterion of translation candidates obtained with context similarity measures.

Our approach most closely resembles (Koehn and Knight, 2002) in that, just like them, we use identical words as our seed lexicon. The difference is that we iterate the calculation of translation equivalents, extending the seed lexicon on every step with additional information, such as context-checked cognates and translation equivalents of most frequent words in the corpus. We also carry out a final cognate-based reranking of translation candidates similar to (Saralegi et al., 2008).

As opposed to (Koehn and Knight, 2002), we are working with much larger corpora and much closer languages, which is why our seed lexicon is much larger, yielding a higher recall as well as precision of the extracted translation equivalents that consequently results in a more usable resource in a real-world setting. And finally, we are not limiting our experiments to nouns, but are working with all content words.

3 Building Resources

In this section we present two resources we built for this experiment: the comparable corpus and the seed lexicon. Since our goal in the experiment reported in this paper is the extraction of translation equivalents for the general vocabulary, we built a Croatian-Slovene comparable news corpus from the 1 billion-word hrWaC and the 380 million-word slWaC that were constructed from the web by crawling the .hr and .si domains (Ljubešić and Erjavec, 2011). We extracted all documents from the domains jutranji.hr and delo.si, which are on-line editions of national daily newspapers with a high circulation and a similar target audience. The documents were already tokenized, PoS-tagged and lemmatized, resulting in 13.4 million tokens for Croatian and 15.8 million tokens for Slovene.

Unlike many language combinations with English, no machine-readable dictionary is available for Croatian and Slovene. Having said this, it is also true that Croatian and Slovene are very close languages. Namely, according to (Scannell, 2007), the cosine for 3-grams in Croatian and Slovene of is 74%, compared to only 34% for English and German that (Koehn and Knight, 2002) used, while a similar result as for Croatian and Slovene was obtained for Czech and Slovak (70%) and Spanish and Portuguese (76%). This means that the lack of dictionary resources for such language pairs can be compensated by exploiting the similarities between the languages. We therefore decided to build a seed lexicon from the comparable news corpus by extracting all identical lemmas that were tagged with the same part of speech in both languages.

As Table 1 shows, the seed lexicon contains about 33,500 entries, 77% of which are nouns. Manual evaluation of 100 random entries for each part of speech shows that nouns perform the best (88%) and that the average precision of the lexicon for all parts of speech is 84%.

The errors we observed in manual evaluation are mostly Croatian words that appeared in the Slovene part of the corpus. They probably orig-

PoS	Size	Precision
nouns	25,703	88%
adjectives	4,042	76%
verbs	3,315	69%
adverbs	435	54%
total	33,495	84%

Table 1: Analysis of the seed lexicon.

inated from readers' comments that are written in informal language which often contains Croatian expressions. Such errors could be avoided in the future by a stricter filtering of the corpus. However, more serious problems could be caused by some false friends that got into the seed lexicon (e.g. "*neslužben*" which means "*unofficial*" in Croatian but "*not part of sbd's job*" in Slovene) and should be addressed in our future work.

4 Extracting Translation Equivalents

In the experiment presented in this paper, our task is to extract a bilingual lexicon from a comparable corpus. The seed lexicon we use to translate features of context vectors was compiled automatically and contains words from the corpus which are identical in both languages. The translation equivalents obtained with this seed lexicon represent the baseline which we then try to beat by extending the seed lexicon with cognates and first translation candidates of the most frequent words in the corpus and a final reranking of the translation equivalents based on cognate clues.

4.1 Experimental Setup

Throughout the experiment we use best-performing settings for building and comparing context vectors from our previous research (see (Ljubešić et al., 2011)). We build context vectors for all content words in each language with a minimum frequency of 50 occurrences in the corpus. The co-occurrence window is 7 content words with encoded position of context words in that window, and log-likelihood as association measure. Vector features are then translated with the seed lexicon, after which Jensen-Shannon divergence is used as similarity measure.

Finally, ten top-ranking translation candidates are kept for automatic and manual evaluation. We try to improve the results by extending the seed lexicon with contextually confirmed cognates as well as with first translations of the most frequent words. In addition, we rerank the translation candidates of all content words obtained with this procedure by taking into account cognate clues among the candidates. The details of lexicon extension and reranking are described in the following sections.

4.2 Evaluation Framework

Automatic evaluation and comparison of the results is performed on a gold standard that contains 1000 randomly selected entries of nouns (618), adjectives (217) and verbs (165) from a traditional broad-coverage Croatian-Slovene dictionary which contains around 8,100 entries. Although we include adverbs in seed lexicon extensions based on their positive impact on this task, we do not include them in the gold standard for two reasons: (I) many tokens tagged as adverbs in the corpus are mistagged other parts of speech and (II) most adverbs in both Croatian and Slovene can be easily generated from adjectives and there is only a small amount of those for which this does not hold, and they can be considered a closed word class.

Mean reciprocal rank (Vorhees, 1999) on the ten top-ranking translation candidates is used for calculating precision. In this experimental setup, recall for nouns is always 45% because we always find translations for 278 of the 618 nouns from the gold standard that satisfy the frequency criterion (50) in the source corpus and have at least one translation in the target corpus that meets the same frequency criterion. For other parts of speech recall is also constant: 42% for adjectives and 56.4% for verbs. Overall recall is 46.2%. The baseline precision used for evaluating seed lexicon extensions of 0.592 was calculated by translating features in context vectors of nouns, verbs and adjectives with the seed lexicon of identical words using the settings described in the previous section. Baseline precision for individual parts of speech is 0.605 for nouns, 0.566 for adjectives and 0.579 for verbs. For a more qualitative insight into the results we also performed manual evaluation of each experimental setting on a sample of 100 random translation equivalents.

4.3 Extending the Seed Lexicon with Cognates

In order to beat the baseline we first extended the seed lexicon with cognates. We calculated them with BI-SIM, the longest common subsequence of

bigrams with a space prefix added to the beginning of each word in order to punish the differences at the beginning of the words (Kondrak and Dorr, 2004). The threshold for cognates has been empirically set to 0.7.

In this step, translation equivalents were calculated as explained above for all content words (nouns, adjectives, verbs and adverbs), taking into account 20 top-ranking translations and analyzing them for cognate clues in that order.

If we found a translation equivalent that met the cognate threshold of 0.7, we added that pair to the lexicon. If the seed lexicon already contained a translation for a cognate we identified with this procedure, we replaced the existing lexicon entry with the new identified cognate pair. Replacing entries is a decision based on empirical results.

PoS	Size	Precision
nouns	1,560	84%
adjectives	779	92%
verbs	706	74%
adverbs	114	85%
total	3,159	84%

Table 2: Manual evaluation of cognates.

As Table 2 shows, we identified more than 3,000 contextually proven cognates, almost half of which are nouns. Manual evaluation of 100 random cognates for each part of speech shows that cognate extraction is most accurate for adjectives (92%), probably because of the regular patterns used to form adjectives in Croatian and Slovene (e.g. Cro. "*digital*an", Slo. "*digital*en", Eng. "*digital*").

Manual evaluation shows that the quality of the extracted cognates on all parts of speech but nouns is substantially higher than the quality of identical words used to generate the seed lexicon. These results can be explained by the different extraction methods for identical words and for cognates: while full string matching was the only criterion for extracting identical words, cognates had to meet an additional criterion – they had to appear in similar enough contexts (i.e. among the 20 top-ranking translation candidates calculated with the context similarity measure). Experimenting with a context similarity threshold as well as a minimum frequency criterion for identical words did not improve the results. On the other hand, we use context-dependent cognates because calculating cognates between all lemmata of specific parts of speech proved to be very noisy even on high cognate thresholds and it did not have a positive impact on this task. Nouns have a higher precision on identical words than on contextually proven cognates probably because of a high amount of proper nouns in the corpus.

Table 3 contains the results of automatic evaluation of bilingual lexicon extraction with the seed lexicon that was extended with cognates. Nouns and adjectives contribute to the task the most, although the amount of adjectives added to the lexicon is half the size of nouns. Adding all parts of speech to the lexicon improves the results for 0.061.

When taking into account specific parts of speech, nouns experience the biggest improvement (0.103) while, interestingly, adjectives show a decrease in precision. " Adjectives, however, show the biggest improvement if only nouns are added to the seed lexicon. The reason for that is probably the syntactic similarity of Croatian and Slovene because of which, since we encode the position in features as well, adjectives are precisely matched between languages if primarily nouns co-occuring with them are taken into account. A similar, but less strong improvement can be observed with verbs that obtain the highest results if only cognate adverbs are added to the seed lexicon.

lexicon	N	A	V	all
baseline	0.605	0.566	0.579	0.592
cognates-N	0.657	0.578	0.596	0.630
cognates-Adj	0.669	0.567	0.590	0.634
cognates-V	0.630	0.497	0.555	0.589
cognates-Adv	0.604	0.573	0.608	0.598
cognates-all	0.708	0.534	0.604	0.653

Table 3: Automatic evaluation of translation extraction with a seed lexicon including cognates.

4.4 Extending the Seed Lexicon with First Translations of the Most Frequent Words

We have shown that precision of the first translation candidates of highly frequent words in the corpus is especially high (Fišer et al., 2011). We therefore decided to add them to the seed lexicon as well and see if they can improve the quality of the task of bilingual lexicon extraction. We only took into account the first translation candidates

for words that appear in the corpus at least 200 times. If the seed lexicon already contained an entry we were able to translate with this procedure, we again replaced the old pair with the new one.

PoS	Size	Precision	Cognates
nouns	2,510	71%	48%
adjectives	957	57%	38%
verbs	1,002	63%	30%
adverbs	325	59%	26%
total	4,794	62%	34%

Table 4: Manual evaluation of first translations of the most frequent words.

Overall, first translation candidates yielded 1,635 more entries for the seed lexicon than cognates but their quality is much lower (by 22% on average). More than 52% of the extracted first translation candidates are nouns, which are also of the highest quality (71%) according to manual evaluation performed on a random sample of 100 first translation equivalents for each part of speech. It is interesting that many of the manually evaluated first translation candidates were also cognates, especially among nouns (48%), further strengthening the argument for using cognates in bilingual lexicon extraction tasks for closely related languages. In 23% of the cases the incorrect translation candidates were semantically closely related words, such as hypernyms, co-hyponyms or opposites that are not correct themselves but probably still contribute to good modeling of contexts and thereby help bilingual lexicon extraction.

Table 5 gives the results of automatic evaluation of bilingual lexicon extraction with the seed lexicon that was extended with first translation candidates. As with cognates, nominal first translations have the most impact on the size of the extended lexicon (2,510 new entries), but share an almost identical precision gain with adjectives. Best performance, again, is achieved when adding all parts of speech to the seed lexicon improving the baseline results by 0.113, 85% more than in case of adding cognates to the seed lexicon. This shows a higher importance of adding high-frequency first translation candidates to the seed lexicon as opposed to adding contextually proven cognates.

When analyzing the precision on specific parts of speech, nouns again experience the largest precision increase of 0.152 (a 48% increase when compared to cognates). The situation with adjectives resembles the one observed when cognates were added to the seed lexicon. This time, adding all parts of speech did not decrease precision, but again, the highest precision is obtained when adding only first translation nouns to the seed lexicon (a 141% higher increase than when adding all parts of speech). This shows once again the importance and potential simplicity of adding syntactic information to the task by just weighting parts of speech on specific positions differently when extracting a specific part of speech.

lexicon	N	A	V	all
baseline	0.605	0.566	0.579	0.592
first-N	0.665	0.665	0.626	0.659
first-Adj	0.700	0.581	0.589	0.656
first-V	0.643	0.513	0.546	0.599
first-Adv	0.610	0.583	0.581	0.599
first-all	0.757	0.607	0.639	0.705

Table 5: Automatic evaluation of translation extraction with a seed lexicon including first translations.

4.5 Combining Cognates and First Translations of the Most Frequent Words to Extend the Seed Lexicon

In order to study the total impact of seed lexicon extension with new information that was extracted from the corpus automatically, we combine the cognates and first translation candidates in order to measure the gain of both information sources. Thereby the seed lexicon was extended with 2,303 new entries, amounting to 35,798 entries overall. When we start adding cognates and then add first translations of most frequent words (overwriting the existing lexicon entries with new information), we achieve precision of 0.731 while changing the order gives a slightly lower score of 0.723. This shows once again that first translations are more beneficial for the context vector translation for bilingual lexicon extraction.

Manual evaluation of a random sample of 100 translation equivalents we extracted from the best-performing extended seed lexicon shows that 88 entries contained the correct translation among the ten top-ranking translation candidates and that 64 of those were found in the first position while 24 were found in the remaining nine positions. This significantly outperforms our baseline of 0.592.

What is more, many lists of ten top-ranking

translation candidates contained not one but several correct translation variants. Also, as many as 59 of correct translation candidates were cognates and 41 of them appeared in the first position, suggesting that the results could be improved even more by a final reranking of translation candidates based on cognate clues which we describe in the following section.

4.6 Reranking of Translation Candidates with Cognate Clues

Once we obtained translation candidates ranked according to our similarity measure, the final reranking of 10 highest-ranking translation candidates was performed. The source word was compared by the previously described BI-SIM function with each of the ten translation candidates. Two lists were formed, one with words satisfying the 0.7 cognate threshold, and another one with the words not satisfying the criterion. Finally, the lists were merged by putting the cognate list of translation equivalents in front of the non-cognate list.

PoS	Baseline	Extended	Reranking
nouns	0.605	0.768	0.848
adjectives	0.566	0.605	0.698
verbs	0.579	0.658	0.735
all	0.592	0.713	0.797

Table 6: Automatic evaluation of translation extraction per part of speech with reranking.

Table 6 shows the baseline results for all parts of speech, the results obtained by using the extended seed lexicon, and the results of reranking the final translation candidates. As expected, the biggest gain through reranking is achieved for adjectives (15.4%), probably because of the regularity of patterns for forming adjectives in both languages. Nouns and verbs experience a similar precision boost (around 11%).

Regarding the final results, the best score is achieved for nouns with a total precision increase of 40%. Although adjectives experience the biggest boost by reranking, their extraction precision is still the lowest. The observations made about their sensitivity to parts of speech being encoded in their context vectors should therefore be exploited in further research. The overall improvement of the results for all parts of speech is 34.6%.

These figures confirm the positive impact of exploiting language similarity on knowledge-light extraction of bilingual lexicons from comparable corpora for closely related languages. Last but not least, the described method results in a fully automatically created resource the quality of which already makes it a useful resource for practical tasks.

5 Conclusions and Future Work

In this paper we presented a knowledge-light approach to bilingual lexicon extraction from comparable corpora of similar languages. When tested on a comparable news corpus for Croatian and Slovene, it outperforms related approaches both in terms of precision (0.797 for nouns, adjectives and verbs) and recall (46%). Unlike most related approaches it deals with all content words not just nouns, and enriches the seed lexicon used for translating context vectors from the results of the translation procedure itself, thereby experiencing a 35% precision increase on the lexicon extraction task. The proposed approach is directly applicable on a number of other similar language pairs for which there is a lack of bilingual lexica.

In the future, we plan to extend our approach to multi-word expressions as well because they are an important component for most HLT tasks. We plan to exploit the observed positive impact of preferring specific parts of speech when calculating translation equivalents of other parts of speech. Additionally, we wish to address polysemy by refining the translation procedure of context vectors as well as measuring similarity of contexts within and across languages.

Acknowledgments

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Programme (FP7/2007-2013), grant agreement no. 248347, and by the Slovenian Research Agency, grant no. Z6-3668.

References

Al-Onaizan, Y. and Knight, K. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. In: *ACL'02*, pp. 400-408.

Fišer, D., Ljubešić, N., Vintar, Š and Pollak, S. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In: *BUCC'11*.

Fung, P. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In: *AMTA'98*, pp. 1-17.

Ismail, A. and Manandhar, S. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In: *COLING'10*, pp. 481-489.

Koehn, P. and Knight, K. 2002. Learning a translation lexicon from monolingual corpora. In: *ULA'02*, pp. 9-16.

Kondrak, G. and Dorr, B. J. 2004. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In: *COLING'04*.

Ljubešić, N., Fišer, D., Vintar, Š and Pollak, S. Bilingual Lexicon Extraction from Comparable Corpora: A Comparative Study. In: *WOLER'11*.

Ljubešić, N. and Erjavec, T. 2011. Compiling web corpora for Croatian and Slovene. In: *BSNLP'11*.

Markó, K., Schulz, S. and Hahn, U. 2005. Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons. In: *RANLP'05*, pp. 301-307.

Och, F. J. and Ney, H. 2000. Improved Statistical Alignment Models. In: *ACL'00*, pp. 440-447.

Otero, P. G. and Campos J. R. P. 2005. An Approach to Acquire Word Translations from Non-parallel Texts. In: *EPIA'05*, pp. 600-610.

Otero, P. G. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In: *MTS'07*, pp. 191-198.

Rapp, R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In: *ACL '99*, pp. 519-526.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. 2008. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In: *BUCC'08*.

Scannell, K. P. 2007. Language Similarity Table http://borel.slu.edu/crubadan/table.html.

Shao, L. and Ng, H. T. 2004. Mining New Word Translations from Comparable Corpora. In: *COLING'04*.

Shezaf, D. and Rappoport, A. 2010. Bilingual Lexicon Generation Using Non-Aligned Signatures. In: *ACL'10*, pp. 98-107.

Vorhees, E.M. 1999. TREC-8 Question Answering Track Report. In: *TREC-8*, pp. 77-82.

Xiao, Z. and McEnery, A. 2006. Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. In: *Applied Linguistics* 27(1): 103-129.