Bootstrapping Bilingual Lexicons from Comparable Corpora for Closely Related Languages

Nikola Ljubešić[1] and Darja Fišer[2]

[1] Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr
[2] Faculty of Arts, University of Ljubljana,
Aškerčeva 2, 1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Abstract. In this paper we present an approach to bootstrap a Croatian-Slovene bilingual lexicon from comparable news corpora from scratch, without relying on any external bilingual knowledge resource. Instead of using a dictionary to translate context vectors, we build a seed lexicon from identical words in both languages and extend it with context-based cognates and translation candidates of the most frequent words. By enlarging the seed dictionary for only 7% we were able to improve the baseline precision from 0.597 to 0.731 on the mean reciprocal rank for the ten top-ranking translation candidates with a 50.4% recall on the gold standard of 500 entries.

Keywords: bilingual lexicon extraction, cognates, comparable corpora.

1 Introduction

Bilingual lexicons are indispensable in most cross-lingual NLP applications and their compilation remains a major bottleneck in computational linguistics. Techniques for automatic extraction of translation equivalents from parallel texts have become well established [9] but since parallel corpora are scarce resources, especially for uncommon language pairs and domains, they often cannot be used. This is why an alternative approach has gained momentum in the past decade that relies on texts in two languages which are not parallel but comparable [2], [12] and therefore more readily available, especially from the increasingly rich web data [17].

The approach relies on the assumption that the term and its translation appear in similar contexts [2], [12], which means that a translation equivalent of a source word can be found by identifying a target word with 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 translate the features of source context vectors and compute similarity measures on those. At this point we seem to be caught in a vicious cycle: the very reason why we are resorting to a highly complex comparable corpus approach for mining translation equivalents is the fact that we do not have a bilingual dictionary to use in the first place. This issue has largely remained unaddressed in previous research, which is why we propose a knowledge-light approach

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 91–98, 2011.

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 next section we give an overview of previous 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

Most research into bilingual lexicon extraction from non-parallel texts was inspired by [2] and [12] whose main assumption is that the term and its translation share similar contexts. The method consists of two steps: modeling of contexts and measuring similarity between the source-language and target-language contexts using a dictionary. The majority of approaches follow the bag-of-words paradigm and represent contexts as weighted collections of words using LL [3], TF-IDF [2] or PMI [16]. After word contexts have been built 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 [2], Jaccard [10] or Dice [11].

Central to comparing context vectors across languages is the translation of features in context vectors, 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 [6] show that it is possible to obtain a seed dictionary from identical and similarly spelled words. Slightly differently, [1] and [15] take advantage of transliteration rules for Arabic/Chinese to generate translation candidates, which is especially efficient for named entities and new vocabulary. At the subword level, [8] constructed string substitution rules to obtain cognates in Spanish and Portugese. As an addition to the standard approach, [13] use string similarity as a reranking criterion of translation candidates obtained with context similarity measures.

Our approach is closest to [6] in that we too use identical words as our seed dictionary with the difference that we iteratively extend the seed dictionary on every step and, since we are working with more similar languages, our extracted lexicon is of a higher quality and therefore more usable in a real-world setting.

3 Resources Used

3.1 Building a Comparable Corpus

In this experiment, we wish to extract translation equivalents for the general vocabulary. This is why 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 [4]. 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.

3.2 Building a Seed Dictionary

Since no open-source machine-readable dictionary is available for Croatian and Slovene, we built a seed dictionary from the comparable news corpus by extracting all identical lemmas tagged with the same part of speech in both languages. With this, we exploit the strong similarity between Croatian and Slovene.[1] As Table 1 shows, the seed dictionary contains almost 33,500 entries, 77% of which are nouns. Manual evaluation of 100 random entries for each PoS shows that average precision of the dictionary is 72%, nouns performing the best (88%). The errors are mostly Croatian words found in the Slovene part of the corpus, probably originating from readers' comments which could be avoided by filtering the corpus by language on a sub-document level. There are also some false friends that probably cause more serious problems (e.g. *neslužben* which means *unofficial* in Croatian but *not part of sbd's job* in Slovene).

Table 1. Size and precision of the seed dictionary

POS	Size	Precision
nouns	25,703	88%
adjectives	4,042	76%
verbs	3,315	69%
adverbs	435	54%
total	33,495	72%

3.3 Building a Gold Standard

For automatic evaluation and comparison of the results we built a small gold standard that contains 500 randomly selected nominal entries from a traditional broad-coverage Croatian-Slovene dictionary.

4 Experimental Setup

The goal of this experiment is to extract a bilingual lexicon from a comparable corpus with a seed dictionary of words from the corpus that are identical in both languages. We consider the translation equivalents obtained with this seed dictionary as baseline and then try to improve the results by extending the seed dictionary with contextually confirmed cognates and first translation candidates of the most frequent words. Throughout the experiment we are using best-performing settings for building and comparing context vectors as confirmed by our previous research [5]. Context vectors are built for all content words that appear in the corpus at least 50 times. The co-occurrence window is 7 content words, with encoded position of context words in that window, and Log-Likelihood as vector association measure. Vector features are then translated with the

[1] According to [14] the cosine for 3-grams in Croatian and Slovene of 74%. A similar similarity was computed for Czech-Slovak (70%) and Spanish-Portuguese (76%).

seed dictionary, after which Jensen-Shannon Divergence is used as a vector similarity measure. Finally, ten top-ranking translation candidates are kept for automatic and manual evaluation.

The evaluation measure is mean reciprocal rank. Although we extract translations for all content words, we report here the results of the automatic evaluation for nouns only due to space restrictions. In this experimental setup, recall is always 50.4% because we always find translations for 252 of the 500 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. To calculate the baseline, we translated features in context vectors with the seed dictionary of identical words. Using the settings described above we achieve 0.597 precision for the baseline.

4.1 Adding Cognates to the Seed Dictionary

In this step of the experiment we augment the seed dictionary with cognates. They are calculated with BI-SIM [7], a modified bigram longest common subsequence function. The threshold for cognates has been empirically set to 0.7. First, translation equivalents are calculated as explained above taking into account 20 top-ranking translations. If we find a translation equivalent that meets the cognate threshold, we add that pair to the dictionary. We test dictionary expansion in two ways: by overwriting the existing dictionary entry with the identified cognate pair and by leaving the existing dictionary entry and disregarding the identified cognate pair.

Table 2. Manual evaluation of cognates

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

As Table 2 shows, we identified more than 3,000 cognates, almost half of which are nouns. Manual evaluation of 100 random cognates for each part of speech shows that cognate extraction performs best 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*).

It is interesting to see that the quality of the extracted cognates is 12% higher than the quality of the identical words. The reason for this is probably the contextual verification of cognates.

Table 3 contains the results of automatic evaluation of bilingual lexicon extraction with the seed dictionary that was augmented with cognates. Overwriting the existing dictionary entries with the new translation always performs better than leaving the old translation. By augmenting the seed dictionary with cognates, a 0.088 increase in precision is achieved.

Table 3. Automatic evaluation of bilingual lexicon extraction using the seed dictionary augmented with cognates (OW: existing entries were overwritten with cognate pairs, NOW: existing entries were kept)

POS	Size	New	Precision-OW	Precision-NOW
baseline	33,495	0	0.597	0.597
cognates-N	34,089	1,560	0.626	0.612
cognates-Adj	33,999	779	0.657	0.639
cognages-V	33,655	706	0.621	0.613
cognates-Adv	33,565	114	0.598	0.598
cognates-NAdj	34,593	2,339	0.679	0.641
cognates-all	34,823	3,159	0.685	0.649

4.2 Adding First Translation Candidates to the Seed Dictionary

In our previous research we showed that the precision of the first translation candidates of highly frequent words in the corpus was especially high [5]. We therefore decided to add to the seed dictionary the first translation candidates for words that appear in the corpus at least 200 times. If the seed dictionary already contains an entry, we again test dictionary expansion in the same two ways as described above.

Overall, first translation candidates yielded 1,635 more entries for the seed dictionary than cognates but their quality is much lower (by 21.5% on average). Almost 53% of the extracted first translation candidates are nouns, which are of the highest quality (71%) according to manual evaluation performed on a random sample of 100 first translation equivalents for each PoS. It is interesting to note that many of the manually evaluated first translation candidates were also cognates, especially among nouns (48%), which further strengthens the argument for using cognates in bilingual lexicon extraction tasks. The incorrect translation candidates were in 22.5% of the cases 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 helping bilingual lexicon extraction.

Table 5 gives the results of automatic evaluation of bilingual lexicon extraction with the seed dictionary that was augmented with first translation candidates. Again, overwriting the existing dictionary entries with the new translation outperforms leaving the old translation.

Table 4. Manual evaluation of first translation candidates for high-frequent words

POS	Size	Precision	Cognates	Related
nouns	2,510	71%	48%	9%
adjectives	957	57%	38%	9%
verbs	1,002	63%	30%	2%
adverbs	325	59%	26%	4%
total	4,794	62.5%	35.5%	6%

Table 5. Automatic evaluation of bilingual lexicon extraction using the seed dictionary augmented with first translation candidates (OW: existing entries were overwritten with the extracted translation pairs, NOW: existing entries were kept)

POS	Size	New	Precision-OW	Precision-NOW
baseline	33,495	0	0.597	0.597
1st_trans-N	33,964	2,510	0.662	0.625
1st_trans-Adj	33,967	957	0.652	0.620
1st_trans-V	33,695	1,002	0.641	0.609
1st_trans-Adv	33,818	325	0.611	0.598
1st_trans-NAdj	34,436	3,467	0.711	0.650
1st_trans-all	34,817	4,794	0.714	0.651

When first translation candidates of all four PoS are added to the dictionary, precision is 0.117 over the baseline, outperforming cognates by 0.029. This suggests that first translation candidates of most frequent words have a greater impact on translating context vectors and on the quality of the extracted bilingual lexicon.

4.3 Combining Cognates and First Translation Candidates to Extend the Seed Dictionary

Finally, we combine the cognates and first translation candidates in order to measure the information gain obtained by applying both methods simultaneously. Since overwriting existing dictionary entries with new translation pairs consistently achieved better results than keeping the old ones, we only evaluate the former setting here. An additional goal of this experiment is to check which information is more beneficial for extracting translation equivalents from a comparable corpus without an external dictionary, cognates or first translation candidates. This is why in one version of the seed dictionary cognates were added first and then first translation candidates (enabling cognates to be overwritten by translation equivalents) while the second version was built the other way around (enabling translation equivalents to be overwritten by cognates).

When we prefer first translation candidates over cognates, we achieve precision of 73.1% while changing the preference gives a slightly lower score of 72.3%. This shows that first translations are more beneficial for the context vector translation procedure even when this information is combined.

Manual evaluation of a random sample of 100 translation equivalents we extracted with the best-performing augmented seed dictionary shows that 88 contained the correct translation among the ten top-ranking translation candidates. In the first position 64 of those were found and 24 in the remaining nine positions. 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, suggesting that the results could be improved even more by a final re-ranking of translation candidates based on cognate clues.

5 Conclusions and Future Work

In this paper we presented a knowledge-light approach to bilingual lexicon extraction from comparable corpora of similar languages. It outperforms related approaches both in terms of precision (0.731) and recall (50.4%). Unlike most related approaches it deals with all content words, and enriches the seed dictionary used for translating context vectors from the results of the translation procedure itself. The proposed approach is directly applicable on a number of other similar language pairs for which there is a lack bilingual lexicons due to socio-economic reasons.

In the future, we wish to refine the methods for building the comparable corpus. We are also looking into possibilities to extend the approach in such a way that it will be able to handle multi-word expressions as well because they are an important component for most HLT tasks. And, last but not least, 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

1. Al-Onaizan, Y., Knight, K.: Translating Named Entities Using Monolingual and Bilingual Resources. In: ACL 2002, pp. 400–408 (2002)
2. Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: Farwell, D., Gerber, L., Hovy, E. (eds.) AMTA 1998. LNCS (LNAI), vol. 1529, pp. 1–17. Springer, Heidelberg (1998)
3. Ismail, A., Manandhar, S.: Bilingual lexicon extraction from comparable corpora using in-domain terms. In: COLING 2010, pp. 481–489 (2010)
4. Ljubešić, N., Erjavec, T.: hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. In: Proceedings of the 3rd International Workshop on Balto-Slavonic Natural Language Processing, Plze, Czech Republic, (September 1–5, 2011)
5. Fišer, D., Ljubešić, N., Vintar, Š., Pollak, S.: Building and using comparable corpora for domain-specific bilingual lexicon extraction. In: Proceedings of the 4th ACL-HLT Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, Portland, Oregon, USA, (June 24, 2011)
6. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: ULA 2002, pp. 9–16 (2002)
7. Kondrak, G., Dorr, B.J.: Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In: COLING 2004 (2004)
8. Markó, K., Schulz, S., Hahn, U.: Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons In: RANLP 2005, pp. 301–307 (2005)
9. Och, F.J., Ney, H.: Improved Statistical Alignment Models. In: ACL 2000, pp. 440–447 (2000)
10. Otero, P.G., Campos, J.R.P.: An Approach to Acquire Word Translations from Non-parallel Texts. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 600–610. Springer, Heidelberg (2005)

11. Otero, P.G.: Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In: MTS 2007, pp. 191–198 (2007)
12. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: ACL 1999, pp. 519–526 (1999)
13. Saralegi, X., San Vicente, I., Gurrutxaga, A.: Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In: BUCC 2008 (2008)
14. Scannell, K.P.: Language similarity table, http://borel.slu.edu/crubadan/table.html
15. Shao, L., Ng, H.T.: Mining New Word Translations from Comparable Corpora. In: COLING 2004 (2004)
16. Shezaf, D., Rappoport, A.: Bilingual Lexicon Generation Using Non-Aligned Signatures. In: ACL 2010 pp. 98–107 (2010)
17. Xiao, Z., McEnery, A.: Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. Applied Linguistics 27(1), 103–129 (2006)