Translation inference across dictionaries via a combination of graph-based methods and co-occurrence statistics

Thomas Proisl, Philipp Heinrich, Stefan Evert, Besim Kabashi

Corpus Linguistics Group, Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract. This system description explains how to use several bilingual dictionaries and aligned corpora in order to create translation candidates for novel language pairs. It proposes (1) a graph-based approach which does not depend on cyclical translations and (2) a combination of this method with a collocation-based model using the multilingually aligned Europarl corpus.

1 Introduction

Translation of lexical items is a fundamental problem in computational linguistics which plays an important role not only in machine translation, but also in various more specific tasks such as mapping of queries, tags, denotators, and alike across different languages. With ever more bilingual lexicons being electronically available for some language pairs, the problem arises of how to use them to create new bilingual dictionaries.

The organizers of the shared task on Translation Inference Across Dictionaries (TIAD) provided partial bilingual dictionaries for the following four language chains for the eight languages German (*de*), English (*en*), Portuguese (*pt*), Japanese (*ja*), Spanish (*es*), Dutch (*nl*), Danish (*da*), and French (*fr*):

1. German–English–Portuguese
2. German–Japanese–Spanish–Portuguese
3. German–Danish–French–Spanish–Portuguese
4. German–Dutch–Spanish–Danish–French–Portuguese

The resulting language graph is visualized in Figure 1. In addition, the four chains also include Portuguese–German dictionaries for "closing the loop" (dashed edge). According to the task guidelines, use of the Portuguese–German dictionaries is limited to validation purposes. The objective of the task is to create three new dictionaries (dotted edges): German–Portuguese, Danish–Spanish and Dutch–French.

A naïve approach to that problem would be to recursively collect all translation candidates: For each source word, take all translations of that word from the source–pivot$_1$ dictionary; then, for each translation, take all translations from the pivot$_1$–pivot$_2$ dictionary and so on until the target language is reached.

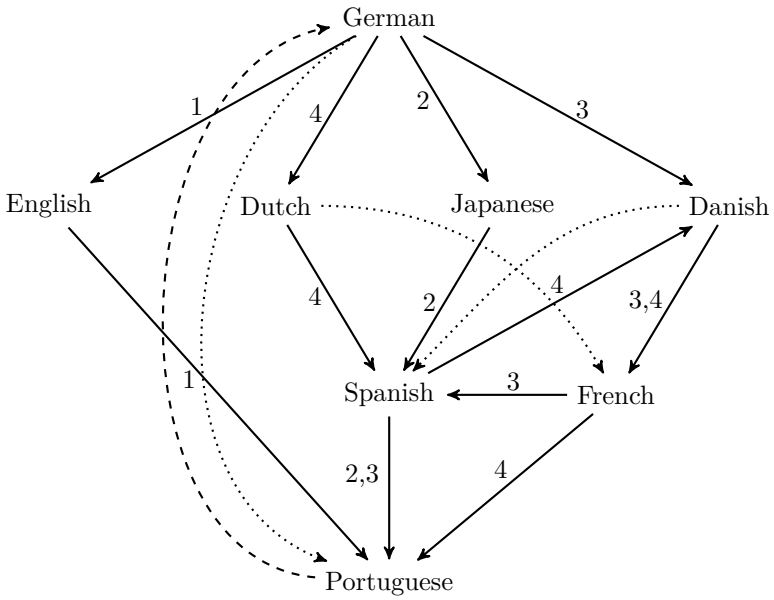

Fig. 1. The language graph. Numbers on the edges show which language chains in the above enumeration are using the respective edge. Dotted edges indicate the desired new direct translation paths.

The problem with this approach is that it results in very noisy and divergent dictionaries.

A common solution to that problem is to make use of cycles (cf. Section 2), in this case by utilizing the Portuguese–German dictionaries. We opted for a novel approach: Instead of relying on cycles, we apply a weighting scheme. We also experiment with combining the translation candidates found via this graph-based approach with candidates extracted from parallel corpora.[1]

2 Related work

The automatic creation of multilingual dictionaries, especially the macro-structure of their entries and annotation interfaces (Kernerman, 2011) as well as the exploitation of resources such as aligned corpora and existing bilingual dictionaries, have attracted commercial and academic research projects for obvious reasons.

Tanaka and Umemura (1994), for example, construct a bilingual dictionary using a third language as a pivot language by utilizing the structure of dictionaries and the lexical entries (nouns). They measure the nearness of the meaning of the

[1] We are talking about *candidates*, since automatic translation techniques yield n-best lists of terms. Both the evaluation function which ranks the candidate terms as well as the precise value of n are at the very core of lexical translation research.

lexical entries to distinguish between true translation equivalents and spurious ones introduced as a result of ambiguity in the pivot language. Similarly, Kaji et al. (2008) construct a Japanese–Chinese dictionary using English as intermediate language. They use monolingual corpora of the first and second language to eliminate the spurious translations caused by the ambiguity of the third language. The wide-coverage monolingual corpora provide the basis for extracting word associations in one language and translation candidates in the target language. This method enables generating domain-specific translation candidates.

Villegas et al. (2016) infer new translations for the languages in a graph of as many as 22 bilingual dictionaries. They consider translation candidates up to three languages away and assign a confidence score to those candidates, which is based on the density of cycles containing the potential target. A cycle is a translation chain which starts and ends at the same lexical item (for a formal definition of translation chains, see section 3.1). Similarly, Mausam et al. (2009) rely on cycles ("translation circuits" in their terms) to match senses probabilistically, and Saralegi et al. (2011) improve precision in pivot-based automatic creation of bilingual dictionaries by inverse consultation, i. e. by looking up translation candidates for all the possible candidates in the target language in the source language. This, however, only works if dictionaries in both directions are at disposal.

Haghighi et al. (2008), on the other hand, do not use a third language at all: They learn bilingual dictionaries only using monolingual corpora and word features in each language. Last but not least, using noisy dictionaries as input, Shezaf and Rappoport (2010) present a method for generating higher-quality dictionaries: their method requires two (noisy) bilingual dictionaries (from the source language to the target language and vice versa) and two comparable monolingual corpora (one in the source language and one in the target language) as input and calculate similarity scores for translation candidates based on the number of words co-occurring with the source word that can be translated into words co-occurring with the target word.

The collocation-based approach described in the present paper, on the other hand, employs a similar idea as can be found in Kovář et al. (2016), who use a transformation of the Dice coefficient for extracting translation candidates from parallel corpora with sentence alignment.

3 System description

3.1 Graph-based approach

As mentioned in Section 1, we opted for a novel graph-based approach that does not rely on cycles. Instead, we use a weighting scheme. As an additional, self-imposed constraint, we do not make use of the Portuguese–German dictionaries at all.

For our weighting scheme, we do not only use the four paths provided by the task organizers but all available simple chains from a source language to

a target language. Simple chains are paths that ignore the orientation of the edges and where no vertex can occur twice. We distinguish between language chains, i. e. chains from one language to another, as illustrated in Figure 1, and translation chains, i. e. chains from one word to another, via the languages in a given language chain.

Formally, let $L_{s,t}$ denote the set of language chains from source language s to target language t. Each language chain $\ell \in L_{s,t}$ is assigned a weight

$$w_\ell = \frac{1}{(|\ell| + |r_\ell|)}, \tag{1}$$

where $|\ell|$ is the length of the chain and $|r_\ell|$ is the number of edges in ℓ that are traversed in reverse. The weights are normalized such that

$$\sum_{\ell \in L_{s,t}} w_\ell = 1. \tag{2}$$

The intuition behind these weights is that the more intervening languages we have and the more dictionaries we use in reverse, the more the quality suffers. Therefore, short chains should get a higher weight than long ones and using a dictionary in reverse should be penalized.

Let $R_{w,\ell}$ denote the set of translation chains from word w in the source language of a language chain ℓ to words in the target language of that language chain. Each translation chain $r \in R_{w,\ell}$ connects w to a potential translation equivalent $e = \tau(r)$. Each translation equivalent e in the set of translation equivalents

$$E_{w,\ell} = \{\tau(r) | r \in R_{w,\ell}\} \tag{3}$$

is assigned a weight

$$w_{e,\ell} = \frac{|\{r \in R_{w,\ell} | \tau(r) = e\}|}{|R_{w,\ell}|}. \tag{4}$$

This weight corresponds to the relative frequency of translation chains from w to e via the languages in language chain ℓ.

Now that we have weights for all language chains and for all translations along a language chain, we can obtain all translation equivalents in the target language t for word w from the source language s, i. e. $E_w = \bigcup_{\ell \in L_{s,t}} E_{w,\ell}$. Each translation equivalent $e \in E_w$ is assigned a weight

$$w_e = \sum_{\ell \in L_{s,t}} w_\ell w_{e,\ell}. \tag{5}$$

The weights are normalized such that $\sum_{e \in E_w} w_e = 1$.

Now we can simply select the n translation equivalents with the highest weights. But what is a suitable value for n, i. e. how can we determine the best number of translation equivalents for a given word? Let $R_w = \bigcup_{\ell \in L_{s,t}} R_{w,\ell}$ be

the set of all chains from word w in the source language s to words in the target language t. Then, we set

$$n = \left\lceil |E_w|^{\frac{1}{c}} \right\rceil, \tag{6}$$

where $\lceil x \rceil$ is the ceiling function and $c = \sum_{r \in R_w} |r| / |R_w|$. This means we approximate n by the average number of translations for each word along the translation chains for word w.

3.2 Collocation-based approach

We make use of the Europarl corpus (see Koehn, 2005: release v7) in its preprocessed and sentence-aligned form (Tiedemann, 2012)[2]. As a further preprocessing step, all monolingual corpora except for the Portuguese one are lemmatized with off-the-shelf algorithms. Unfortunately, we did not lemmatize the Portuguese corpus in time. For the language pair de–pt, our procedure thus yields lexical surface realizations as translation candidates (see below).

We retrieve translation candidates by analyzing first-order (syntagmatic) collocations. The procedure is implemented via the R-package `wordspace` (Evert, 2014)[3]. For each language pair, lemmata (or, in the case of Portuguese, types) are extracted together with their alignment beads from the corpus in order to create lemma-sentence matrices with the intersection of alignment beads as columns. As an example, the French corpus contains 28,100 lemmata, the Dutch one 36,048, and there is an intersection of 2,003,463 alignment beads.

These matrices are then transformed into one term-term co-occurrence matrix for each language pair. The nl–fr co-occurrence matrix from the example above has thus 36,048 rows and 28,100 columns. Subsequently, the Dice score is calculated for each lemma of the source language (if it occurs in the corpus) and each target term. The Dice score is a de-facto standard for the determination of translation candidates (Smadja et al., 1996) and represents the harmonic mean of the conditional probabilities $\mathbb{P}\{\text{source}|\text{target}\}$ and $\mathbb{P}\{\text{target}|\text{source}\}$. Let O_{11} denote the co-occurrence frequency of source and target term, R_1 the marginal frequency of the target term and C_1 the marginal frequency of the source term (notation and formula taken from Evert, 2008), then the Dice score can be calculated by means of

$$\text{dice}\,(O_{11}, R_1, C_1) = \frac{2O_{11}}{R_1 + C_1}. \tag{7}$$

The higher its value, the higher the association between source and target term. Thus, for every source term, the target terms with the highest Dice scores serve as translation candidates. Note that in this step we ignore all candidates which solely consist of punctuation marks and/or digits in order to improve translation quality.

[2] http://opus.lingfil.uu.se/Europarl.php
[3] http://wordspace.r-forge.r-project.org/

3.3 Combination of collocation-based and graph-based approaches

Without having an evaluation measure which determines the trade-off between precision and recall of the translation candidates, we opted for a very simple combination of the two approaches above: the final list of candidates is gained by union of the graph-based candidates and four collocation-based candidates.[4]

4 Evaluation

The evaluation procedure was announced after submission of the translation candidates and solely takes precision (and no recall)[5] into account. For each language pair and system, 100 source-target-candidates were sampled. Subsequently, each translation pair was reviewed manually according to whether the target term was a correct (possible) translation of the source term.

Two scalar performance measures are given, see Table 1: Precision is the percentage of (manually determined) correct translations among the proposed candidates. Additionally, "gold-precision" only labels those candidates as "true positives" which can be found in the organizers' (undisclosed) gold-standard of translations.

5 Results and discussion

Results for the graph-based approached ("graph") and the combination of collocation-based and graph-based approaches ("combined") can be found in Table 1. Two findings seem noteworthy: Firstly, the solely graph-based method consistently outperforms the combined approach for both evaluation measures. Secondly, in the nl–fr language-pair setting, both systems are drastically beaten by the baseline, whereas in the other two settings both systems outperform the baseline.

Obviously, our strategy of providing multiple translation candidates proved to be suboptimal for the official task evaluation, which only focused on precision. Note however that our system is easily adaptable in case a reasonable evaluation measure is given a priori: both graph-based and collocation-based methods yield n-best lists of candidates with a scalar score-function enabling a more sophisticated selection of actual candidates.[6]

[4] The graph-based method yields between two and three candidates on average depending on the language pair. Assuming an overlap of one or two candidates between both methods, this heuristic guarantees that the collocation-based approach delivers approximately two additional candidates.

[5] Note that recall is not well-defined in the case of lexical translation: while human experts may easily agree on some unambiguous translations (thus making it feasible to create a gold-standard for calculating precision), they might disagree quickly on particular or unusual translations (thus making it impossible to create a gold-standard for measuring recall).

[6] That is to say: if the system is to focus on precision, a very small number of candidates should be given, and their selection should be based on the distribution of the score functions of both the graph-based and the collocation-based candidate lists.

	precision		gold-precision		
pair	graph	combined	graph	combined	baseline
da–es	**0.93**	0.73	0.35	0.25	0.59
nl–fr	0.44	0.37	0.28	0.25	**0.52**
de–pt	**0.85**	0.72	0.19	0.16	0.62

Table 1. Evaluation measures for all language pairs for both submitted systems (based on samples of size 100). Precision is the percentage of correct translations among the sampled candidates, gold-precision is the percentage of correct translations that were also part in the organizers' gold standard. The baseline figures were provided by the task organizers and are based on a depth-first search for cycles of translations which include the desired source and target languages.

Advantages of our proposed graph-based system are twofold: Firstly, it does not require cycles, i. e. it can be applied in greater variety of settings. Secondly, the weighting scheme takes into account the number of dictionaries involved and the directionality in which they are used on the one hand, and, on the other, the relative frequency of translation chains leading to a translation equivalent; thus, the system automatically determines a suitable number of translation equivalents.

The proposal of further candidates retrieved from the Europarl corpus has turned out to be counterproductive for the reasons elaborated above. Nevertheless, given more realistic settings in which recall of all (or most) possible translations is important, retrieval of candidates not comprised in any of the bilingual corpora (or of those with atypical translation paths) seems desirable. Future work will thus use more sophisticated methods for combining graph-based and collocation-based candidates, e. g. by using the Borda count or the Schulz method.

Bibliography

Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:1212–1248, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.6220&rep=rep1&type=pdf.

Stefan Evert. Distributional Semantics in R with the wordspace Package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 2014. URL http://aclweb.org/anthology/C14-2024.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2008, pages 771–779, 2008. URL http://www.aclweb.org/anthology/P08-1#page=815.

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. Automatic Construction of a Japanese-Chinese Dictionary via English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 699–706, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/175_paper.pdf.

Ilan Kernerman. From dictionary to database: Creating a global multi-language series. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10–12 November 2011*, pages 113–21, 2011. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.698.6117&rep=rep1&type=pdf.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.5497&rep=rep1&type=pdf.

Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*, 29(3):339–352, September 2016. ISSN 0950-3846, 1477-4577. doi: 10.1093/ijl/ecw029. URL https://doi.org/10.1093/ijl/ecw029.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 262–270. Association for Computational Linguistics, 2009. URL http://aclweb.org/anthology/P09-1030.

Xabier Saralegi, Iker Manterola, and Inaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011. URL http://aclweb.org/anthology/D11-1078.

Daphna Shezaf and Ari Rappoport. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association*

for Computational Linguistics, pages 98–107. Association for Computational Linguistics, 2010. URL http://aclweb.org/anthology/P10-1011.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996. URL http://aclweb.org/anthology/J96-1001.

Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics – Volume 1*, pages 297–303. Association for Computational Linguistics, 1994. URL http://aclweb.org/anthology/C94-1048.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Marta Villegas, Maite Melero, N. Bel, and J. Gracia. Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 868–876, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/613_Paper.pdf.