Auto-generating bilingual dictionaries: Results of the TIAD-2017 shared task baseline algorithm

Morris Alper

K Dictionaries, Tel Aviv, Israel
E-mail: morris@kdictionaries.com

Abstract

Inferring a bilingual dictionary L1>L3 given two bilingual dictionaries L1>L2 and L2>L3 is a non-trivial task, as seen in reports of large-scale, computationally-heavy experiments published in recent years (Soderland et al. (2009); Shezaf and Rappoport (2010)). Early works on this (cf. Tanaka and Umemura (1994)) have already noticed that the main obstacle in such inferences stems from the fact that polysemy is not isomorphic across languages, and often a monosemous lexical item in L1 can be polysemous in L2. In this paper, we present a new set of experiments on automatically generating bilingual dictionaries based on existing ones. The data used are a commercial set of bilingual dictionaries with a particular topology when viewed as a graph connecting source and target languages. We find that searching for cycles in this graph is an effective method for generating translation inferences, and reflect on the impact of the source data's structure on these results and directions for future research.
Keywords: automatic dictionary generation, bilingual lexicography, polysemy

1. Introduction

With high-performance, full-fledged machine translation systems such as Google Translate and Bing Translator, the idea of generating bilingual dictionaries may seem to be a relatively easy task. After all, translating sentences (let alone larger textual units) is seemingly a much harder task than generating translation candidates for a lexical item. Consider the last sentence, which contains 23 words: disambiguating all the words in context and transferring them to a well-formed sentence in another language involves many NLP components beyond the lexical item level, including those dealing with lexical and structural ambiguity, word order, anaphora resolution, and so forth. Most of these are not relevant when suggesting equivalents for a word like 'task' taken from the sentence and viewed in isolation.

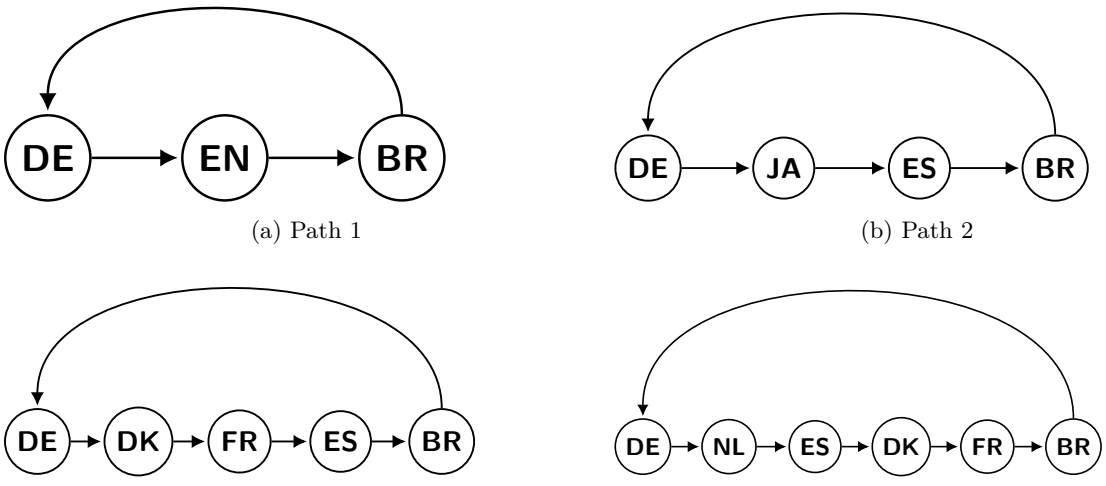
Bilingual dictionaries, however, are composed not only of translation equivalents for source-language lexical items. Compiling a bilingual dictionary requires selecting the lexical items that are deemed worthy of inclusion, providing morphological information for the given translations, introducing usage examples, deciding on the order of the translations, providing glosses for lexical gaps, and more. Note that the latter three tasks require meta-linguistic knowledge and even a certain theory of meaning representation which is far beyond the reach of current models of statistical and neural machine translation systems. Virtually all systems that auto-generate bilingual dictionaries are restricted to producing bilingual lexicons. It turns out that even this is a non-trivial task due to what has been called *anisomorphism*: "[W]hile it is possible to find translation equivalents at the sentence level, it is more difficult at the level of lexical units. This difficulty has its origin in the cultural component which exists in every language and which causes words, which are dynamic and explicit symbols of that culture, not to have full and absolute equivalents in other languages. This fact strongly affects some fields of knowledge; for example business and economics, because they tend to be closely related to particular cultures." (al Qāsimī et al. 1977, as cited in Fuertes-Olivera and Arribas-Baño 2008)

Our methodology is computationally straightforward: the algorithm starts with L1 and goes to L2 then L3 (and L4, L5, etc.), and ends with a translation from the last language in the chain back to L1. By starting with a given sense in L1 and finally retrieving it again as a translation in the last pair of the chain, we reinforce the confidence in our selection. These chains correspond to cycles in the graph of lexical items as vertices connected by edges when a translation is present. In general we expect that such cycles occur when meaning is preserved across translations, so that the same sense is recovered once returning to the original language, and thus we can infer a translation between non-adjacent pairs of lexical items in the chain.

The main contribution of our method as regards previous research is an analysis of the problem definition and graph structure of the source data on the resulting translation inferences. We consider the contributions of which languages are connected in the dataset, the directedness of translations, and language typology. We also discuss the methodology of evaluating the performance of such translation inference systems and we provide directions for further research that take advantage of the a broader range of available lexicographic data.

2. Dataset

K Dictionaries (KD) possesses rich lexicographic resources for various languages, compiled using a standard format. In this experiment we used a subset of the data contained in KD's bilingual dictionaries. We generated translation inferences for pairs of languages for which KD already possesses traditional bilingual dictionaries, which can be used to expedite the evaluation of automatically generated translations.



(a) Path 1 (b) Path 2

(c) Path 3 (d) Path 4

Fig. 1: Language paths in the dataset

We included four language paths in our dataset, as shown in Figure 1. Each of these begins with the same set of 100 randomly-selected words in German, and each successive language in the path includes translations of the words from the previous language. For example, the German noun *hässlich* translates to *ugly*, *mean*, and *nasty* in English. Each of these in turn has multiple translations into Brazilian Portuguese, which will then have multiple translations into German.

The same language may be reached in different ways (e.g. Spanish is reached via the paths DE>JA>ES, DE>DK>FR>ES, and DE>NL>ES) and will contain non-identical sets of words depending on how it is reached. In total we arrive at 2279 German headwords as the final translations, summing the contributions from the four paths, which illustrates the exponential growth of number of translations in bilingual dictionaries that are recursively chained.

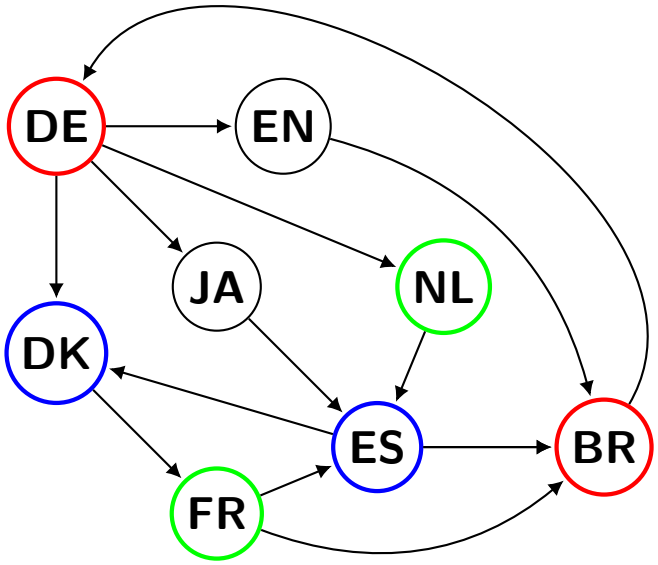


Fig. 2: Dictionary topology, with three target language pairs (DE>BR, DK>ES, NL>FR) color-coded

We can combine these paths together into a graph representing the topology of our dataset, as shown in Figure 2. Note that by design the graph is connected and cyclic (i.e. contains cycles). Although the given graph is directed, we will later discuss the extent to which translations may be treated as reflexive.

3. Goal

Three language pairs were selected as targets for our translation inference algorithm: German>Brazilian Portuguese (DE>BR), Danish>Spanish (DK>ES), and Dutch>French (NL>FR). Although we did not provide translations from the first to the second member of each pair in the given dataset, we do have these translations in KD's dictionaries which can be used to partially verify automatically-generated translations.

Note that there are edges connecting BR>DE and ES>DK, while our goal is to produce edges in the opposite direction connecting these nodes. We additionally include the restriction that the BR>DE edge cannot be directly reversed, though we allow reversing the ES>DK edge.

4. Algorithm

Our algorithm consists of finding cycles of translations in the graph shown in Figure 2. The idea is that although translations diverge as pivot languages are traversed, we can increase the confidence in recursive translations if we arrive at the same headword we started from upon returning to the source language. For example, consider the following translation path:

DE	**DK**	**FR**	**ES**	**BR**	**DE**
darstellen	*fremstille*	*fabriquer*	*fabricar*	*fabricar*	*herstellen*
"represent"	"manufacture"	"fabricate"	"fabricate"	"fabricate"	"produce"

Here there has been some semantic shift across successive translations, due to non-overlapping semantic ambiguity of lexical items in these different languages. As a result, we arrive at a different word in German at the end of the translation path, and so it may be discarded. By contrast, consider the following translation path:

DE	**NL**	**ES**	**BR**	**DE**
darstellen	*weergeven*	*representar*	*representar*	*darstellen*
"represent"	"represent"	"represent"	"represent"	"represent"

In this case the meaning has remained approximately constant along the translation path, and we return to the same word in German at which we began.

Although our graph is directed, we generalize our algorithm using the simplifying assumption that translations are reflexive, so if say English *apple* translates to French *pomme* then we may assume that French *pomme* can translate to English *apple*. Thus we elect to search for translation cycles irrespective of the given translation directions, except for the restriction given above that we may not reverse the BR>DE edge. This allows us to find more translation cycles such as the following:

DK	**ES**	**NL**	**DE**	**DK**
se på	*mirar*	*bekijken*	*betrachten*	*se på*
"watch (v.)"	"watch (v.)"	"view (v.)"	"view (v.)"	"watch (v.)"

As can be seen in Figure 2, the edges DK<ES, ES<NL, NL<DE, and DE>DK do not form a directed cycle, but they do form an undirected cycle, and we expect that these will also correspond to relatively accurate translations.

The complete graph consisting of our dataset and all translations consists of thousands of densely-connected nodes and finding all cycles in this graph would be computationally infeasible. However, this issue is obviated by including the reasonable restriction that translation paths may only traverse any given language once. Then we can find cycles efficiently with a depth-first search from the word whose translation is desired.

To summarize, our algorithm consists of the following: For each pair of languages for which we must generate translations, we have source language L_s and target language L_t. For each node (headword) n_s in L_s, we perform a depth-first search beginning at n_s for cycles, with the following constraints:

- the search is undirected except for the edge BR>DE which may only be traversed in that direction
- cycle may contain at most one node in each language
- cycle must contain a node in L_t

This terminates once we have found such a cycle, and we infer that the node in L_t is the translation of the node in L_s.

5. Results

For each language pair, we measured the number of lexical items in the source language ("Source nodes") for which we would like to find translations, and the number of translations inferred by the algorithm ("Inferences"). We also estimated the precision of our algorithm by sampling the inferred translations and manually checking their accuracy; we checked all inferred translations for the pairs DE>BR and NL>FR, and a random sample of 100 out of the 251 inferred translations for the pair DK>ES. The column "Estimated precision" contains the ratio of the number of sampled translation inferences that were judged to be correct translations to the total number of translation inferences sampled. The column "Gold precision" contains the ratio of the number of translation inferences that were found in KD's gold standard dictionaries containing direct translations between these language pairs over the total number of translation inferences.

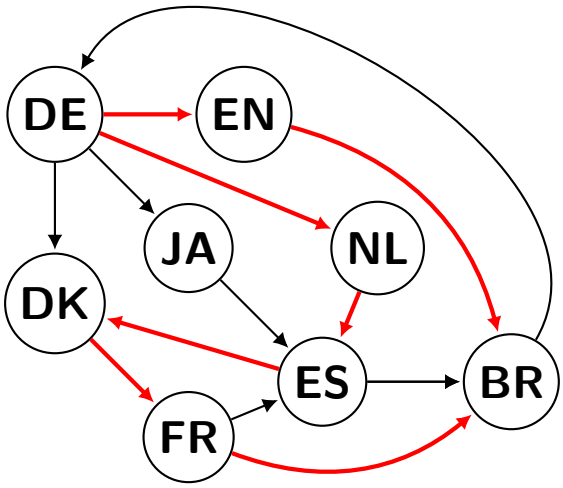
Language pair	**Source nodes**	**Inferences**	**Estimated precision**	**Gold precision**
DE > BR	100	50	1.00	0.62
DK > ES	1039	240	0.89	0.59
NL > FR	145	63	0.70	0.52

Note that we have not calculated recall because it is not well-defined for this task. Recall would measure the fraction of all possible translations that have been discovered, but existing bilingual dictionaries do not purport to be exhaustive lists of all possible translations for each headword and thus this cannot be reliably measured.

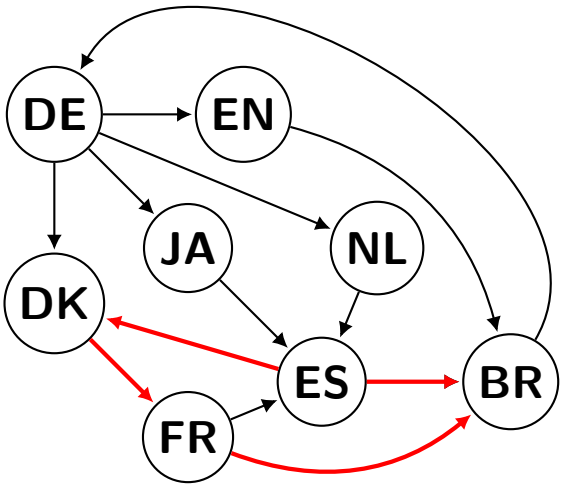
We also examined the number of occurrences of different cycle shapes within the graph of languages. Figure 3 lists the four most common paths through the dictionary. Paths (a) and (b) both require some translations to be reversed, while (c) and (d) are traversable while respecting directedness.

6. Discussion

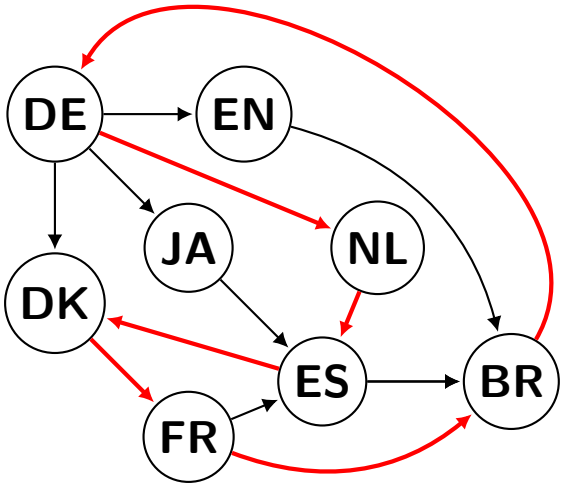
Our algorithm produces translation inferences with reasonable accuracy ($\geq 70\%$, based on the manually-evaluated precision estimates). While there is room for improvement, these results demonstrate that this simple and computationally inexpensive algorithm could be used to greatly reduce the manual work required to generate a bilingual dictionary in a new pair of languages.



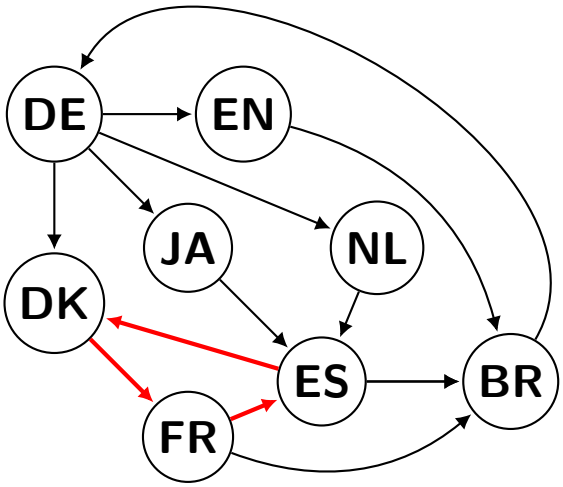
(a) 79 occurences (note reversed paths)



(b) 76 occurences (note reversed path)



(c) 56 occurences



(d) 30 occurences

Fig. 3: Four most common paths through the dictionary

The precision estimates calculated from the gold standard data are quite different than the precision estimated based on manual evaluation of sampled inferences, which implies that the gold standard dictionaries are far from being exhaustive with respect to the translations provided for each headword. Indeed, it is reasonable to assume that the goal of a bilingual dictionary is not necessarily to provide every possible translation equivalent, and the selection of translations provided in such a dictionary is the product of many factors including editorial choice and differing tolerance for semantic deviation. This relates to the difficulty in measuring recall in order to evaluate such translation inference algorithms; it remains for future research to examine how to effectively measure the extent to which an algorithm's inferred translations provide sufficient coverage.

Note that the performance of the algorithm is quite different for different language pairs. With highest precision for the pair DE>BR and lowest for NL>FR, it is apparent that the algorithm is affected by considerations of language typology and/or the graph structure of the dataset. Since the most common paths all contain both Romance and Germanic languages together, the contribution of language typology does not conform to the expectation that better translations should be inferred from chains of languages from the same family (though the lack of Japanese in these cycles might be related to it being the typological outlier among the languages in the dataset). Regarding graph structure, in the given dataset the languages DE and BR (and similarly DK and ES) are connected while the shortest path between NL and FR crosses another node. This matches the intuition that better translations can be produced between pairs of languages that are more closely connected in the dataset. Since these results were obtained using a restricted subset of the language translation pairs present in KD's lexicographic resources, we expect that these results will improve when the algorithm is run using all available data.

Recall that the algorithm treated the graph as undirected in order to find cycles of translations. The assumption that translations are reflexive is not generally valid with regards to dictionaries. For example, in the dataset one of the English translations listed for the German word 'Abitur' is 'high school graduation'. While this is an accurate translation, the phrase 'high school graduation' is not a lexical unit that would normally appear as a headword in a dictionary of English. Similarly, translations may consist of inflected forms which should not occur as headwords. However, such forms will not have translations themselves and are less likely to appear as translations of headwords from other languages, so this is less significant for our cycle-based algorithm. On the other hand, allowing reversed translations in paths does significantly affect the number of cycles that can be found, as evidenced by the fact that the two most common cycles found in the language graph were undirected (see Figure 3).

Among various limitations of the current algorithm is that it only selects one translation for each lexical item in the source language, present in the first cycle found in the depth-first search. This conceivably could be improved upon by finding all cycles containing the source item and a lexical item in the target language, and selecting lexical items in the target language in such cycles that satisfy some measure of goodness (e.g. cycle length) as translations. In addition, the current algorithm does not use much of the rich lexicographic data which is present in the source resources including synonyms and antonyms, semantic fields, example sentences, and other data. We hypothesize that these components could be used to increase the level of confidence of existing translations and remove invalid translations from consideration.

7. Conclusion

We have presented a computationally straightforward method for automatically generating bilingual dictionaries based on existing ones. By finding cycles of translations in the graph of all lexical entries with translations treated as undirected edges, we were able to infer translations with reasonable accuracy. We found that precision is best estimated using manual evaluation rather than searching existing dictionaries for the inferred translations, and an examination of the most common paths traversed by the algorithm implied that treating the translations as undirected was significant in the algorithm's performance. Future research will focus on how to measure the exhaustiveness of the translations

produced, how to effectively compare multiple cycles containing the same source lexical item, and how to use the supporting lexicographic data connected to dictionary headwords to improve the quality of generated translations.

Bibliography

al Qāsimī, A. M. et al. (1977). *Linguistics and bilingual dictionaries*. Brill Archive.

Fuertes-Olivera, P. A. and Arribas-Baño, A. (2008). *Pedagogical Specialised Lexicography: The representation of meaning in English and Spanish business dictionaries*, volume 11. John Benjamins Publishing.

Shezaf, D. and Rappoport, A. (2010). 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.

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., et al. (2009). 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-Volume 1*, pages 262–270. Association for Computational Linguistics.

Tanaka, K. and Umemura, K. (1994). 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.