Auto-generating Bilingual Dictionaries

Noam Ordan[1], Jorge Gracia[2], Ilan Kernerman[1]

[1] K Dictionaries Ltd, 8 Nahum Hanavi Street, 6350310 Tel Aviv, Israel

[2] Universidad Politécnica de Madrid, Campus de Montegancedo sn, Boadilla del Monte 28660 Madrid, Spain

E-mail: noam@kdictionaries.com, jgracia@fi.upm.es, ilan@kdictionaries.com

Abstract

Inferring a bilingual dictionary from two or more existing bilingual dictionaries is a non-trivial task, as seen in reports of large-scale, computationally-heavy experiments published in recent years. Early works on this have already noted 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 one language can be polysemous in its corresponding translation into another language. In this paper, we propose an experiment on translation inference across dictionaries, based on a graph-based view of a collection of bilingual dictionaries. The idea is to explore the results and analyze them from a lexicographic point of view to reflect the implications that the issue of anisomorphy introduces in the task, and to illustrate its hurdles and potential benefits.

Keywords: automatic dictionary generation, bilingual lexicography, polysemy, translation

1. Introduction

Dictionaries are a human effort at representing meaning, whether of a language on its own terms, or of one language (L1) vis-à-vis another language (L2). Whereas the task of the translator (human or machine) is to find an ad-hoc solution for substituting the meaning of an L1 item with its equivalent in L2 given some context, this solution may be partial or rare or just good enough for the context at hand, but completely useless otherwise, since estimating and reusing such rare events is not realistic within state-of-the-art machine translation systems. Similarly, a bilingual dictionary that would list all the items in L2 that were ever given as equivalents for a unit in L1 would be an impractical resource, even from the point of view of machines, searching over all possible options is considered NP-complete, i.e., unrealistic computationally (Knight, 1999).

Scaling up the human effort required for compiling bilingual dictionaries into a highly multilingual landscape is an inherently difficult task, due to the combinatorial explosion of pair-wise language comparisons. To alleviate this issue, the automatic generation of bilingual/multilingual dictionaries, based on already existent ones, is a research and practical avenue which merits exploration, with the aim of assisting and complementing human-based dictionary compilation.

Our methodology in the current experiment 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 (which we call "closing the loop"), we reinforce the confidence in our selection. In addition, if the loop is not closed in the last chain, we consult another bilingual dictionary (see below).

The rest of the paper is organized as follows: Section 2 describes the data we utilize in our experiments. In Section 3 we present the experiment and in Section 4 the results of the automatically generated translations. Section 5 reports on an additional contribution of our methods, which allows for automatic generation of synonymous and semantically related words, and Section 6 reviews relevant literature, both practical implementations of solutions to the problem and the lexicographic and lexicological obstacles which should be overcome. In Section 7, we conclude with a brief discussion.

2. Dataset

The experiments rely on two subsets of data of K Dictionaries, namely MLDS and KMT:[1]

- MLDS

The Multi-Language Dictionary Series (aka Global Series, cf. Kernerman, 2015), currently contains lexicographic cores for 24 languages. Each consists of approximately 12,000 main entries featuring detailed semantic and grammatical information, including alternative script, word categorization and inflected forms, definitions and examples of usage, word sense disambiguators and various attributes (e.g., synonyms and antonyms, register, sense qualifier), multiword expressions, etc. Several languages have a second level that doubles their size (to about 25,000 entries), and Spanish is quadrupled (50,000 entries). The L1 cores are created from scratch with the idea of minute lexical mapping, and can be used to produce monolingual dictionaries, but serve mainly as a base for integrating translation equivalents and developing bilingual sets. So far, nearly one hundred pairs were developed manually, though their division among L1 cores is unequal: on the one hand, three have no bilingual versions yet, whereas French, on the other hand, is the most extensively translated (into 18 languages). The bilingual versions of a single language core are juxtaposed together, thus forming a multilingual dataset of that L1.

[1] The initial experiments (called Cross-Lingual Automated Common Senses, CLACS) began in-house in 2016, making use of full cross-lingual lexicographic resources. In 2017, the shared task on Translation Inference Across Dictionaries (TIAD) was launched, making available to researchers limited bilingual dictionary resources, with results presented at a workshop held as part of the first Language, Data, Knowledge conference (https://tiad2017.wordpress.com/).

- KMT

The K Multilingual Translators (KMT) (aka MultiGloss, cf. Egorova, 2015; Kernerman, 2015) consist of semi-automatically generated by-products of the English Multilingual Dictionary (KEMD) that include translations in 45 languages. Twenty-two of these translation languages have been reversed and manually edited and refined into detailed bilingual word-to-sense L1 indices to English. Then, the KEMD translations in all the other languages are added to the English equivalents of these bilingual indices, thus producing the multilingual index – linking each sense of the L1 headword via its English counterpart to all the other language translations that are available in KEMD.

3. Procedure

Our graph is rather simple, and we traverse it the following way:

The new bilingual dictionary was generated by using four language pairs from MLDS, as follows:

1. German to Turkish (DE>TR)
2. Turkish to French (TR>FR)
3. French to Brazilian Portuguese (henceforth 'Portuguese', FR>BR)
4. Portuguese (back) to German (BR>DE)[2]

The results were processed with the help of two factors:

1. Check translations from the existing MLDS set from Portuguese to German (i.e., pair 4 above). If a German translation is recognized, we consider it a 'closed' loop – since we begin from a specific sense of a German entry and end up with the same entry.
2. If not found, check for a translation (in 4) in the KMT Portuguese-German resource. Recall that KMT is created semi-automatically, so in terms of confidence we trust more the selection of translations stemming from MLDS. However, we use it as another pivot to validate the inferred translations.

[2] Needless to say, we could reverse the last pair, i.e., Brazilian-Portuguese>German to improve results, but we have self-imposed a restriction to avoid this option so that our study is carried out in lab-clean conditions where only cross-dictionary pivoting is considered.

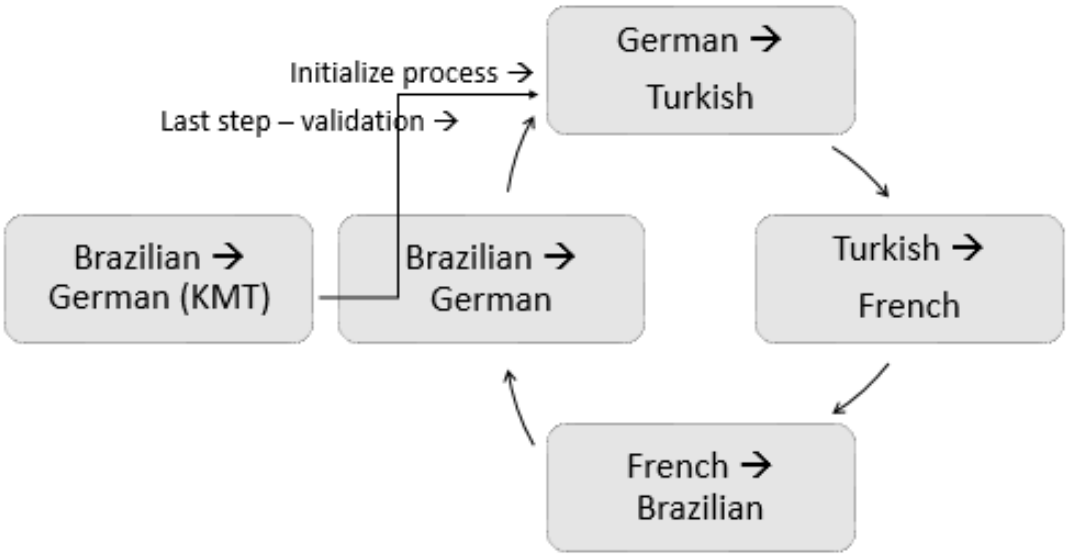

Figure 1: The four language-pair chains with an additional validation step for closing the loop

As we discuss in Section 6, there are many sub-cases of anismorphism across and between languages, and it turns out that more often than not – even though divergence (i.e., one- or few-to-many mapping between L1 and L2) grows exponentially and for each source-language item in German we manage to retrieve a huge number of back-to-German-translations – usually we do not manage to attain the identical German words, although we use two Portuguese>German dictionaries as our final pivot. This is well illustrated in Figure 2. The source-language word *Abkommen* is not found among the eight inferred translations of the last pair that closes the loop, i.e., in the Portuguese to German dictionary.

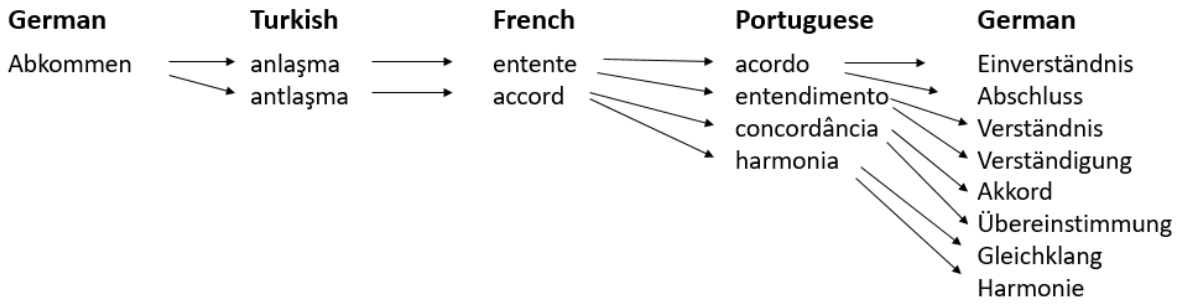

Figure 2: Divergence across languages increasing exponentially
(so the loop does not always close)

4. Findings

We began with 12,000 German entries (from MLDS). The number of entries that were found in both MLDS and KMT (BR>DE) is 5,865. The matches break down according to five quality scores (for the summary statistics see also Table 1). (The total number of matches before closing the loop with KMT was 8,722.)

Quality 1: contains 4,377 entries (74.63%); it is defined operatively as a case-sensitive exact match between the initial headword of the German dictionary plus the part of speech and a translation containing exactly these features (same word, same part of speech) arrived at via a chain of bilingual dictionaries.

Quality 2: contains 44 entries (0.75%); the only difference from the Q1 criterion is a non-match in terms of upper/lower-case letter (which is typical/unique for German).

Quality 3: contains 24 entries (0.41%); the only difference from Q1 and Q2 criteria is the missing article in German, e.g., *Warnung* die *Warnung* (also typical/unique for German).

Quality 4: contains 594 entries (10.13%), where the initial headword is a substring of the final string arrived at in the chain, e.g., *Boden* der *(Erd)Boden*.

Quality 5: contains 826 entries (14.08%), where the initial German headword does not match the final translation arrived at (as regards MLDS), though it does match a translation existing in KMT. For example, *Bestandteil* is potentially synonymous to *Grundbestandteil*. This quality score generates our candidates for synonyms (see Section 5).

quality score	raw frequency	ratio
1	4,377	74.63%
2	44	0.75%
3	24	0.41%
4	594	10.13%
5	826	14.08%
Total	5,865	100.00%

Table 1: Results of automatic matching according to quality

In term of precision, we report ~75% accuracy. This is considerably lower than the 90% accuracy reported in a much more sophisticated algorithm devised in Mausam et al. (2008). However, it should be borne in mind that there are major differences in the setting and the results are not comparable. We suggest that the reasons for the non-

comparability touch on fundamental issues concerning the current task:

1. The number of valid translations for any given word or phrase is much larger than reflected in a bilingual dictionary. Specia and Nunes (2006) estimate that for certain lexical items there are hundreds of possible translations. Our evaluation was done against a medium-sized bilingual dictionary (that had the restriction of offering maximum three translation equivalents), and any item that was automatically inferred and was not found in the dictionary is considered "an error". Mausam et al. (2008), however, sampled hundreds of inferred translations and used crowd-sourcing to decide whether the translations were valid or not. This allows to increase the number of candidates beyond that which is found in a dictionary. Additionally, their evaluation relied on self-proclaimed native speakers, and therefore must be taken with a grain of salt.

2. Given that the number of possible translation is so big, we have no access to the total number of translations for all the entries. This means that recall cannot be calculated, as recall, by definition, is calculated against the total number of relevant/correct items.

5. Synonyms and semantic fields

The lowest Q5 score does not necessarily imply a bad translation, but could indicate potential synonym candidates, or at least semantically-related words of relevance for learners and other users or for computational tasks concerning word-sense disambiguation and information retrieval. We could thus also consider utilizing this architecture to generate *semantic clouds* that surround any word sense in our data.

The intuition behind the generation of these semantic clouds by Q5-scored translations is the following. Consider the scenario we have experimented with: L1>L2>L3>L4>L1. The fourth node, L3>L4, yields a large amount of translations, most of which are noisy. Looking at the resources L4>L1 (MLDS and KMT) is akin to eliciting two judgements, and it stands to reason that if both point back at the same L1 item the chances that the L4 is a good translation candidate for the source-language L1 item increase. However, if one points back at this L1 item (KMT), but the other does not (MLDS), what does it mean? Arguably, and as illustrated in Table 2, both yield valid translations, often synonymous (like *Addresse* and *Anschrifft*).

As we have indicated in Section 3, we preferred to rely more heavily on MLDS as its quality is higher, and therefore preferred to penalize results where MLDS did not have a match and KMT did; however, as can be seen from Table 2, a match in KMT and a non-match in MLDS has three possibilities: (1) a synonym, which can be taken as a valid translation, is yielded; (2) a semantically-related word is retrieved; (3) rarely, a non-related word is retrieved. Our sample space is too small to arrive at statistically meaningful results.

A case of a non-synonymous but closely related words is, for example, *Bestandteil* and *Grundbestandteil*, and it was for this reason that we decided to score differently everything generated through this sub-procedure. Semantically, however, the two words are closely related, meaning *element* and *basic element*, respectively. In other cases, we find that non-matches like these are also related, albeit more vaguely.

Consider the following path:

German		**Turkish**		**French**		**Portuguese**		**German**
Abenteuer	→	*serüven*	→	*aventure*	→	*aventura*	→	*Affäre*

In this case, the match is taken from KMT, so we do have some confidence with respect to the validity of the path/result. The word *Abenteuer* means *adventure*, and *Affäre* means *(love) affair*. This figurative extension indicates a semantic relation.

Headword	**English gloss**	**Synonym candidate**	**English gloss**
Abc	ABC	**Alphabet**	alphabet
Abgrund	precipice, abyss	Welten	worlds
ablehnend	unfavorable	**Negative**	negative
Absatz	paragraph, leap	Sprung	jump
Abschnitt	section	**Absatz**	paragraph
Absicht	intention	**Ziel**	goal
Achtung	danger, esteem	**Wertschätzung**	appreciation
Addresse	address	**Anschrifft**	address
Affe	monkey	Wagenheber	jack (for a car)
Akt	act	**Handlung**	action
autonomy	autonomous	**Unabhängig**	independent
Autonomie	autonomy	**finanzielle Unabhängigkeit**	financial independence
Autoritär	authoritarian	**Diktatorisch**	dictatorial
Backpulver	baking powder	**Hefe**	yeast
Ball	ball (also as in ballroom)	**Fest**	celebration / feast
Bankrott	bankruptcy	**Insolvenz**	insolvency
Barriere	barrier	**Hindernis**	obstacle
Barriere	barrier	**Schranke**	barrier

Table 2: Candidate synonyms generated automatically

Table 2 summarizes 36 more cases for German. As appears below, only three pairs of items are not semantically related, the other 33 (marked in boldface) are either near synonyms or closely related.

6. Related work

Some studies have been reported in the literature. For instance, Tanaka and Umemura (1994) proposed a method to infer indirect translations through a pivot language when constructing bilingual dictionaries. Their pivot is generated by reverse dictionaries and therefore is different than the one suggest here. The loop is thus closed when a two-time inverse consultation is used, such that a set of candidate translations from L1>L2>3 is compared to what they call selection area generated by looking backwards at L3>L2>L1. A modification of their algorithm is suggested by Lim et al. (2011) in the creation of multilingual lexicons.

Other efforts have been done to compile massive multilingual dictionaries automatically, such as Mausam et al. (2008), which rely on the probabilistic exploration of the whole set of translations considered as a graph. Villegas et al. (2016) evolved this notion and moved it to a linked data landscape, applied to the RDF version of the Apertium family of bilingual dictionaries (Gracia et al., 2016). Another method that utilizes corpora and low quality lexicons as seeds is proposed by Shezaf and Rappoport (2010).

These studies illustrate the complexity of the problem, but are hampered by inherent difficulties of the translation process, such as the anisomorphism of languages and lexical gaps. Further, the dictionary compilation process goes far beyond identifying translation candidates. In fact, full-fledged bilingual dictionaries require, among others, selection of L1 lexical items in the first place, division and ordering of their senses, morphological information, usage examples, sense-specific translations in L2 and glosses for lexical gaps, and more. Lexical gaps are just one case out of several that undermines the illusion of a 1:1 mapping between languages; some other cases will be discussed below. As summarized by Adamska-Sałaciak (2006), "a bilingual dictionary cannot perfectly account for meaning, since meaning is always anchored within a particular language"; quoting Zgusta (1971), "the fundamental difficulty of ... a coordination of lexical units [between two languages] is caused by the *anisomorphism* of languages, i.e., by the differences in the organization of designates in the individual languages and by other differences between languages."

Byrd et al. (1987) identify different types of mismatches between languages which violate lexicographic symmetry, notably:

1. Morphologically, some words occur only or mostly as inflected forms, and therefore their lemmatized (uninflected) form should not appear as a translation equivalent. For example, *allege* appears as a lemma in the English to Italian dictionary they used (*Collins English-Italian Dictionary*, CEID), but it is not

provided as a translation in the Italian to English part, since in most cases it is used as a participle.

2. Some languages make specific distinctions not made in another language, and the best way to represent such specific meanings in the target language is with their superordinate equivalents. For example, the French word *fleuve*, which denotes a river that flows into the sea, is normally translated into the more general term *river*, but reversing *river* back to *fleuve* could be a mismatch that should be at least reviewed (it is translatable also to *rivière*, or river that flows into another course of water).

3. In other cases, the opposite case occurs, namely, a more general item is substituted with a more specific one. For example, *book* is translated in CEID into *quaderno* (*notebook*), *bustina* (*of matches*), and *blocchetto* (of tickets). Here the transfer is from a general term to specific ones.

Another form of anisomorphism is lexical gaps.

1. According to Bentivogli and Pianta (2000), "a lexical gap occurs whenever a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words". They, too, use (a digital version of) CEID and estimate that around 7.8 percent English entries are translated to Italian with a free combination of words, that is, there are thousands of English words which exist as concepts in Italian but are not lexicalized. It is interesting to note that the least gapped part-of-speech is nouns (5.6%) and the most gapped one is adverbs (19.3%).

It is clear from the above that such discrepancies between languages increase the more pivots are added to a system, L1>L2>L3>L4, etc. But even using just one pivot language, i.e., inferring L1>L3 from L1>L2 and L2>L3, is no simple automatic procedure.

Previous works indicate that new resources can be created automatically, if not completely from scratch then at least based on existing ones. To do it successfully, rigorous lexicographic practice should be applied side by side with automatic methods, which are more error-prone. Admittedly, large-scale automatic generation which report 90% accuracy (Mausam et al., 2008) seems promising, but it is unlikely that (human) users would want to use dictionaries where every tenth word contains an error.

7. Conclusions

Empirically analysing current and new techniques for automatic inference of translations with the aim of integrating them with the more rigorous lexicographic practice has become a necessity. As a step in this direction, an experiment has been devised to explore the potential and hurdles of the task. An outcome of this experience

has been the creation of benchmark data for the comparison of different translation inference techniques.[3]

Our findings indicate that the growth rate is exponential, and that *closing the loop* is a sound method for higher quality assurance, but using it as a sole method – although highly precise – leads to a relatively low recall. We have plugged in another pivot (KMT) as a second source for closing the loop, and have shown that beyond its role as an extra validation step, it can generate synonyms and semantically-related words.

Our initial experiments are promising, as we obtain relatively high-quality translation results as well as open the ground for automatically generating additional useful by-products like synonyms and semantic fields. These findings serve as a first step for further experimentation with the use of pivots (which, how many, and how) and with incorporating additional components of the entry (subject fields, synonyms, etc.).

In the future, we intend to use the full potential graph of MLDS and plug in KMT in each and every node. The use of multiple ways to close the loop raises further problems, some of which are due to an inflation in the number of suggested translations. Some works report pruning algorithms to resolve the problem (Mausam et al., 2008; Gracia et al., 2016). Theoretically, it would be interesting to study the effect language similarity plays in divergence. It stands to reason that the more similar two languages are, the less divergence one could expect. This, too, calls for a future study.

8. References

Adamska-Sałaciak, A. (2006). *Meaning and the bilingual dictionary: The case of English and Polish*. Peter Lang.

Bentivogli, L. & Pianta, E. (2000). Looking for lexical gaps. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the 9th EURALEX International Congress*. Stuttgart, Germany: Institut für Maschinelle Sprachverarbeitung, pp: 663-669.

Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S. & Rizk, O. A. (1987). Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4), pp. 219-240.

Egorova, K. (2015). Editing an automatically generated index with K Index Editorial Tool. In I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 268-280. Available at: https://elex.link/elex2015/proceedings/eLex_2015_17_Egorova.pdf

[3] Used for the first time in the context of the TIAD-2017 shared task (https://tiad2017.wordpress.com/)

Gracia, J., Villegas, M., Gómez-Pérez, A., & Bel, N. (2016). The apertium bilingual dictionaries on the web of data. *Semantic Web Journal*.

Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 372-383. Available at: https://elex.link/elex2015/proceedings/eLex_2015_24_Kernerman.pdf

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), pp. 607-615.

Lim, L. T., Ranaivo-Malançon, B. & Tang, E. K. (2011). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits* 43, pp. 45-51.

Mausam, Soderland, S., Etzioni, O., Weld, D, Skinner, M. and Bilmes, J. (2008). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Annual Meeting of the Association of Computational Linguistics*. ACL.

Shezaf, D. & Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, 98-107. Stroudsburg, PA: Association for Computational Linguistics.

Specia, L. & Nunes, M. G. V. (2006). Exploiting the translation context for multilingual WSD. In *Text, Speech and Dialogue*. Berlin/Heidelberg: Springer, pp. 269-276.

Tanaka, K. & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th Conference on Computational Linguistics, Volume 1*, pp. 297–303. ACL.

Zgusta, L. (1971). *Manual of Lexicography.* Hague – Paris: Mouton.