Automatic Generation of Translation Dictionaries Using Intermediary Languages

Kisuh Ahn and Matthew Frampton
ICCS, School of Informatics
Edinburgh University
K.Ahn@sms.ed.ac.uk,M.J.E.Frampton@sms.ed.ac.uk

Abstract

We describe a method which uses one or more intermediary languages in order to automatically generate translation dictionaries. Such a method could potentially be used to efficiently create translation dictionaries for language groups which have as yet had little interaction. For any given word in the source language, our method involves first translating into the intermediary language(s), then into the target language, back into the intermediary language(s) and finally back into the source language. The relationship between a word and the number of possible translations in another language is most often 1-to-many, and so at each stage, the number of possible translations grows exponentially. If we arrive back at the same starting point i.e. the same word in the source language, then we hypothesise that the meanings of the words in the chain have not diverged significantly. Hence we backtrack through the link structure to the target language word and accept this as a suitable translation. We have tested our method by using English as an intermediary language to automatically generate a Spanish-to-German dictionary, and the results are encouraging.

1 Introduction

In this paper we describe a method which uses one or more intermediary languages to automatically generate a dictionary to translate from one language, X, to another, Y. The method relies on using dictionaries that can connect X to Y and back to X via the intermediary language(s), e.g. $X \rightarrow IL$, $IL \rightarrow Y$, $Y \rightarrow IL$, $IL \rightarrow X$, where IL is an intermediary language such as English. The resources required to exploit the method are not difficult to find since dictionaries already exist that translate between English and a vast number of other languages. Whereas at present the production of translation dictionaries is manual (e.g. (Serasset1994)), our method is automatic. We believe that projects such as (Boitet *et al.*2002) and (Wiktionary), which are currently generating translation dictionaries by hand could benefit greatly from using our method. Translation dictionaries are useful not only for end-user consumption but also for various multilingual tasks such as cross-language question answering (e.g. (Ahn *et al.*2004)) and information retrieval (e.g. (Argaw *et al.*2004)). We have applied our method to automatically generate a Spanish-to-German dictionary. We chose this language pair because we were able to find an online Spanish-to-German dictionary which could be used to evaluate our result.

The structure of the paper is as follows. In section 2.1, we describe how if we translate a word from a source language into an intermediary language, and then into a target language, the number of possible translations may grow drastically. Some of these translations will be 'better' than others, and in section 2.2 we give a detailed description of our method for identifying these 'better' translations. Having identified the 'better' translations we can then automatically generate a dictionary that translates directly from the source to the target language. In section 3 we describe how we used our method to automatically generate a Spanish-to-German dictionary, and in section 3.3, we evaluate the result. Finally, in section 4, we conclude and suggest future work.

2 Translating Via An Intermediary Language

2.1 The Problem

Consider the problem of finding the different possible translations for a word x from language X in language Y when there is no available $X \rightarrow Y$ dictionary. Let us assume that there are dictionaries which allow us to connect from X to Y and back to X via an intermediary language IL i.e. dictionaries for $X \rightarrow IL$, $IL \rightarrow Y$, $Y \rightarrow IL$ and $IL \rightarrow X$, as shown in figure 1.

If there was only ever 1 suitable translation for any given word in another language, then it would be trivial to use dictionaries $X \rightarrow IL$ and $IL \rightarrow Y$ in order to obtain a translation of x in language Y. However, this is not the case - for any given word x in language X the $X \rightarrow IL$ dictionary will usually give multiple possible translations $(il_1 ... il_n)$, some of which diverge more than others in meaning from x. The $IL \rightarrow Y$ dictionary will then produce multiple possible translations for each of $(il_1 ... il_n)$ to give $(y_1 ... y_p)$ where $p >= n$. Again, some of $(y_1 ... y_p)$ will diverge more than oth-

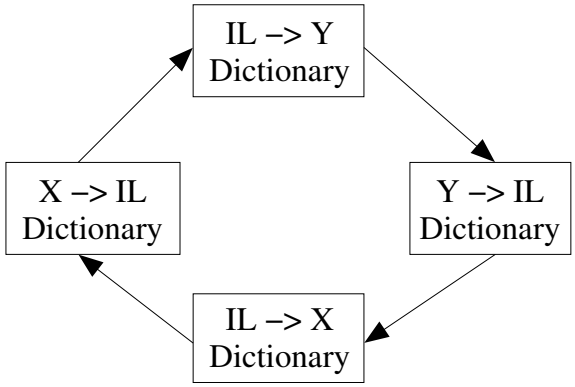

Figure 1: The cycle of dictionaries

ers in meaning from their source words in $(il_1...il_n)$. Hence we have p possible translations of the word x from language X in language Y. Some of $(y_1...y_p)$ will have diverged less in meaning than others from x, and so can be considered 'better' translations. The problem then is how to identify these 'better' translations.

2.2 Using The Link Structure To Find 'Better' Translations

Our method for identifying the 'better' translations is to first use dictionary $Y \rightarrow IL$ to produce $(il2_1...il2_q)$, the multiple possible translations of each of $(y_1...y_m)$, where $q >= m$. Next we use dictionary $IL \rightarrow X$ to give $(x2_1...x2_r)$, the multiple translations of each of $(il2_1...il2_q)$, where $r >= q$. We then select each of the members of the set $(x2_1...x2_r)$ which are equal to the original word x. We hypothesise that to have returned to the same starting word, the meanings of the words that have formed a chain through the link structure cannot have diverged significantly, and so we retrace two steps to the word in $(y_1...y_m)$ and accept this as a suitable translation of x. Figure 2 represents a hypothetical case in which two members of the set $(x2_1...x2_r)$ are equal to the original word x. We retrace our route from these through the links to y_1 and y_2, and we accept these as suitable translations.

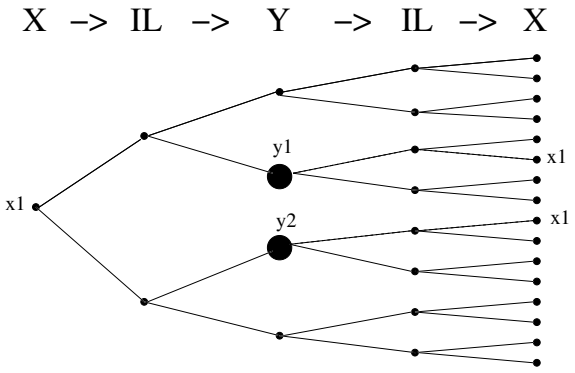

Figure 2: Translating from $X \rightarrow IL \rightarrow Y \rightarrow IL \rightarrow X$. Nodes are possible translations.

If we apply the method described here to a large number of words from language X then we can automatically generate a language X-to-language Y dictionary. Here we have considered using just one intermediary language, but provided we have the dictionaries to complete a cycle from X to Y and back to X, then we can use any number of intermediary languages, e.g. $X \rightarrow IL$, $IL \rightarrow IL2$, $IL2 \rightarrow IL$, $IL \rightarrow Y$, where $IL2$ is a second intermediary language.

3 The Experiment

We have applied the method described in section 2 in order to automatically generate a Spanish-to-German dictionary using Spanish-to-English, English-to-German, German-to-English and English-to-Spanish dictionaries. We chose Spanish and German because we were able to find an online Spanish-to-German dictionary which could be used to evaluate our automatically-generated dictionary.

3.1 Obtaining The Data

We first collected large lists of German and English lemmas from the Celex Database, ((Baayen and Gulikers1995)). We also gathered a short list of Spanish lemmas, all starting with the letter 'a' from the Wiktionary website (Wiktionary) to use as our starting terms. We created our own dictionaries by making use of online dictionaries. In order to obtain the English translations for the German lemmas and vice versa, we queried 'The New English-German Dictionary' site of The Technical Universiy of Dresden [1]. To obtain the English translations for the Spanish lemmas and vice versa, we queried 'The Spanish Dict' website [2]. Finally, we wanted to compare the performance of our automatically-generated Spanish-to-German dictionary with that of a manually-generated Spanish-to-German dictionary, and for this we used a website called 'DIX: Deutsch-Spanisch Woerterbuch' [3]. Table 1 gives information about the four dictionaries which we created in order to automatically generate our Spanish-to-German dictionary. The fifth is the manually-generated dictionary used for evaluation.

Dicts	Ents	Trans	Trans/term
S to E	664	1202	1.8
E to S	18845	28300	1.5
G to E	27918	77626	2.8
E to G	26298	104693	4.0
S to G'	610	3126	5.1

Table 1: Dictionaries; S = Spanish, E = English, G = German, S to G' is the dictionary used for evaluation.

[1] http://www.iee.et.tu-dresden.de/cgi-bin/cgiwrap/wernerr/search.sh

[2] http://www.spanishdict.com/

[3] http://dix.osola.com/

3.2 Automatically Generating The Dictionary

For our experiment, we used the method described in section 2 to automatically construct a scaled-down version of a Spanish-to-German dictionary. It contained 664 Spanish terms, all starting with the letter 'a'. To store and operate on the data, we used the open source database program PostgresSQL, version 7.3.4. Starting with the Spanish-to-English dictionary, at each of stages $1-3$, we produced a new dictionary table with an additional column to the right for the new language. We did this by using the appropriate dictionary to look up the translations for the terms in the old rightmost column, before inserting these translations into a new rightmost column. For example, to create the Spanish-to-English-to-German (SEG) table, we used the English-to-German dictionary to find the translations for the English terms in the Spanish-to-English (SE) table, and then inserted these translations into a new rightmost column. We kept producing new tables in this fashion until we had generated a Spanish-to-English-to-German-to-English-to-Spanish (SEGES) table. In stage 4, the final stage, we selected only those rows in which the starting and ending Spanish terms were the same. Important characteristics of these dictionary tables are given in table 2.

Stages	Dicts	Ents	Trans	Trans/term
0	SE	664	1202	1.8
1	SEG	631	6966	11.0
2	SEGE	568	24012	42.3
3	SEGES	566	38002	67.1
4	SEGES	533	4313	8.1

Table 2: Constructing Dictionary; Ents = number of entries, Trans = number of translations, Trans/term = average number of translations given per entry.

Table 2 shows that the number of translations-per-term grew and grew from 1.8 translations in the starting Spanish-to-English dictionary to an enormous 67.1 translations per term in the SEGES table after stage 3. However, after stage 4, having selected only those rows with matching first and last entries for Spanish, we reduced the number of translations back to 8.1 per term.

3.3 Evaluation

Having automatically generated the Spanish-to-German dictionary containing 533 unique Spanish terms, we then compared it to the manually-generated Spanish-to-German dictionary (see section 3.1). We gave the same initial 664 Spanish terms to the manually-generated dictionary but received translations for only 610.

The results are summarised in table 3. We observe that when we regard the manually-generated dictionary as the Gold-standard, our automatically-generated dictionary managed to produce a relatively adequate coverage of some 73.6% (392 out of 533) with respect

	Auto SG	Man SG	Overlap
Entries	533	610	392 (73.6%)
Total Trans	4313	3126	1577
Trans/Entry	8.1	5.1	4.0 (78.4%)

Table 3: Result: SG automatic vs SG manual

to main entries overlap between the two dictionaries. When we look at the number of translations per term, we find that our dictionary covered most of the translations found in the manually-generated dictionary (4.0 out of 5.1 average or 78.4%) for which there was a corresponding entry in our dictionary. In fact, our dictionary produced more translations-per-term than the manually-generated one. An extra translation may be an error or it may not appear in the manually-generated dictionary because the manually-generated dictionary is too sparse. Further evaluation is required in order to assess how many of the extra translations were errors.

In conclusion, we find that our automatically-generated dictionary has an adequate but not perfect coverage and very good recall for each term covered within our dictionary. As for the precision of the translations found, we need more investigation and perhaps a more complete manually-generated comparison dictionary. The results might have been even better had it not been for several problems with the four starting dictionaries. For example, a translation for a particular word could sometimes not be found as an entry in the next dictionary. This might be because the entry simply wasn't present, or because of different conventions e.g. listing verbs as "to Z" when another simply gives "Z". Another cause was differences in font encoding e.g. with German umlauts. Results might also have improved had the starting dictionaries provided more translations per entry term, and had we used part-of-speech information - this was impossible since not all of the dictionaries listed part-of-speech. All in all given the fact that the quality of data with which we started was far from ideal, we believe that our method shows great promise for saving human labour in the construction of translation dictionaries.

4 Conclusion

In this paper we have described a method using one or more intermediary languages to automatically generate a dictionary to translate from one language, X, to another, Y. The method relies on using dictionaries that can connect X to Y and back to X via the intermediary language(s). We applied the method to automatically generate a Spanish-to-German dictionary, and desptite the limitations of our starting dictionaries, the result seems to be reasonably good. As was stated in section 3.3, we did not evaluate whether translations we generated that were not in the gold-standard manual dictionary were errors or good translations. This is essential future work. We also intend to empirically

test what happens when further intermediary dictionaries are introduced into the chain.

We believe that our method can make a great contribution to the construction of translation dictionaries. Even if a dictionary produced by our method is not considered quite complete or accurate enough for general use, it can serve as a very good starting point, thereby saving a great deal of human labour - human labour that requires a large amount of linguistic expertise. Our method could be used to produce translation dictionaries for relatively unconnected language groups, most likely by using English as an intermediary language. Such translation dictionaries could be important in promoting communication between these language groups and an ever more globalised and interconnected world.

A final point to make regards applying our method more generally outside of the domain of translation dictionary construction. We believe that our method, which makes use of link structures, could be applied in different areas involving graphs.

References

Kisuh Ahn, Beatrix Alex, Johan Bos, Tiphaine Dalmas, Jochen L. Leidner, Matthew B. Smillie, and Bonnie Webber. Cross-lingual question answering with qed. 2004.

Atelach Alemu Argaw, Lars Asker, Richard Coester, and Jussi Kalgren. Dictionary based amharic - english information retrieval. 2004.

R.H. Baayen and L. Gulikers. The celex lexical database (release 2). In *Distriubted by the Linguistic Data Consortium*, 1995.

Christian Boitet, Mathieu Mangeot, and Gilles Serasset. The papillon project: Cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In *2nd Workshop NLPXML*, pages 93–96, Taipei, Taiwan, September 2002.

Gilles Serasset. Interlingual lexical organization for multilingual lexical databases. In *Proceedings of 15th International Conference on Computational Linguistics, COLING-94*, pages 5–9, Aug 1994.

Wiktionary. A wiki based open content dictionary. In *http://www.wiktionary.org/*.