Generating Bilingual Dictionaries by Transitivity

Luka Nerima, Eric Wehrli

Language Technology Laboratory (LATL), Department of Linguistics

University of Geneva

24, rue Général-Dufour, CH-1211 Genève 4, Switzerland

nerima@cui.unige.ch, eric.wehrli@lettres.unige.ch

Abstract

Recently the LATL has undertaken the development of a multilingual translation system based on a symbolic parsing technology and on a transfer-based translation model. A crucial component of the system is the lexical database, notably the bilingual dictionaries containing the information for the lexical transfer from one language to another. As the number of necessary bilingual dictionaries is a quadratic function of the number of languages considered, we will face the problem of getting a large number of dictionaries. In this paper we discuss a solution to derive a bilingual dictionary by transitivity using existing ones and to check the generated translations in a parallel corpus. Our first experiments concerns the generation of two bilingual dictionaries and the quality of the entries are very promising. The number of generated entries could however be improved and we conclude the paper with the possible ways we plan to explore.

1. Introduction

One of the problems faced by MT, that we would like to address, is the need for systems capable of handling not just one pair of languages but several pairs, ultimately a very large number of language pairs. The globalization of commercial and cultural exchanges, for instance over the Internet, creates an ever increasing demand for multilingual linguistic tools and in particular multilingual translation systems. To take another example, in the West-European context, where no less than 15 languages are used for everyday life as well as for administrative, scientific and commercial purposes, the dramatically increasing need for translation tools is by no means satisfied by the existing commercial systems.

The problem of the quadratic growth of the number of translation pairs, n(n–1) for n languages, has often been taken as an argument against transfer-based translation models as opposed to systems based on interlingua (see Boitet, 2001, who argues in favor of Unified Networking Language). In the proposed project, we would like to argue against such a view, and show that the above-mentioned problem can be very effectively restricted.

In the framework of our project, we intend to show that this goal can be achieved (i) by the use of an abstract level of representation -- which abstracts away from several surface-structure cross-linguistic differences (ie. Word order, morphological cases, etc.), and (ii) an object-oriented design, which makes possible the use of generic code while allowing for language-pair specific properties and processes through type extension and redefinition of methods, and (iii) by generating automatically a substantial part of the bilingual dictionaries. This paper focuses on the third point.

2. Multilingual lexical database

The lexical database is composed for each language of (i) a lexicon of words, containing all the inflected forms of the words of the language, (ii) a lexicon of lexemes, containing the syntactic information of the words (corresponding roughly to the entries of a classical dictionary) and (iii) a lexicon of collocations (in fact multi-word expressions, that is collocations and idioms). We call the lexemes and the collocations the *lexical items* of a language.

The multilingual lexical database contains the information for the lexical transfer from one language to another. For storage purposes, we use a relational database management system. For each language pair, the bilingual dictionary is implemented as a relational table containing the associations between lexical items of language A to lexical items of language B. The bilingual dictionary is also bi-directional, i.e. it associates as well lexical items of language B to lexical items of language A. In addition to these links, the table contains transfer information such as translation context, preferences between one to many translations, semantic descriptors, argument matching for predicates (mostly for verbs). The table structures are identical for all pairs of languages.

It is important to mention that although the bilingual lexicon is bidirectional, it is not symmetrical. If a word *v* from language A has only one translation *w* in language B, it doesn't necessarily mean that *w* has only one translation *v*. For instance the word *tongue* is translated in French to *langue*, while in the opposite direction the word *langue* has two translations, *tongue* and *language*. In this case the descriptor attribute from French to English will mention respectively « body part » and « language ». Another element of dissymmetry is preferences between

synonymous translations. For instance the lexicographer can mark his preference to translate *lovely* into the French word *charmant* rather than *agréable*. Of course the opposite translation direction must be considered independently.

What is challenging in this project is that it necessitates as many bilingual tables as the number of language pairs considered, i.e. n(n-1)/2 tables. For instance, for 4 languages, it requires 6 bilingual tables, for 5 languages 10 tables, for 6 languages 15 tables. We consider that an appropriate bilingual coverage (for general purpose translation) requires at least 50'000 correspondences per language pair.

In the frame of this project we consider 5 languages (French, English, German, Italian, Spanish). At the moment we have 4 bilingual dictionaries out of the 10 needed with the number of entries listed in the table below:

	Number of entries
English – French	76'311
German – French	45'492
French – Italian	36'672
Spanish – French	19'226

Table 1: Number of entries per language pair

It is important to mention that all of these 4 bilingual dictionaries were manually created by lexicographers and the quality of the entries is good.

3. Automatic generation

The importance of multilingual lexical resources in MT and, unfortunately, the lack of multilingual lexical resources has motivated many initiatives and research work to establish collaboratively made multilingual lexicons, e.g. the Papillon project (Boitet & al. 2002) or automatically generated multilingual lexicons (see for instance Farwell & al 1992, Aymerich and Cam elo 2007, Gamallo 2007).

In order to achieve the building up of the bilingual database during the duration of the project, we plan to use semi-automatic generation for part of it. For this purpose we will derive, by transitivity, a bilingual lexicon using two existing ones. For instance, if we have bilingual correspondences for language pair A→B and B→C, we can obtain A→C. The generated correspondences will then be validated using a parallel corpus. However, the correspondences that could not be checked this way must be checked manually.

The idea of using a pivot language for deriving bilingual from two existing ones is not new. For instance in (Paik & al. 2004, Ahn & Frampton 2006, Zhang & al. 2007) the reader can find related approaches. The specificity of our experiment is that the starting resources are manually

made, i.e. non noisy, lexicons.

The general schema of the derivation is to use two bilingual lexicons that share the same language and to join their relational tables on the lexical items of the shared language (the pivot language). For instance, if we have a bilingual lexicon for the language pair (English, French) and an other one for the pair (German, French), we can derive a bilingual lexicon for the pair (English, German). The validation of correspondences is also crucial: only those that are of good quality must be kept in the generated lexicon.

More precisely, the process goes as follows:

1) Pick two bilingual tables for language pairs (A, B) and (B, C) and perform a relational natural join. Perform a filtering based on the preference attribute to avoid combinatory explosion of the number of generated candidate correspondences.
2) Validate all the unique correspondences, i.e. the correspondences of lexical items having only one translation. While the lexicon is non symmetrical, this process is performed twice, once for each translation direction. Here we make the hypothesis that the lexicographers did their job well and that if there is only one translation for a word then the word is unambiguous.
3) Validate all the correspondences obtained by a pivot lexical item of type collocation. We consider as very improbable that a collocation is ambiguous.
4) All other correspondences are filtered by two means: a) we retain only those correspondences obtained by correspondences with a preference level greater or equal to the average; b) the retained correspondences are checked in a parallel corpus , i.e. only the correspondences used as translation in the corpus are kept.

The last stage deserves to be described in more detail. First, the parallel corpus is tagged by the Fips tagger (Wehrli, 2007). This has the great advantage to lemmatize the words of the corpus thus allowing to consider lexical items (lexemes and collocations) that are the basic elements of the bilingual correspondences rather than words. This is especially valuable for verbs with particles such as those of English and German. In order to check the validity of the correspondences, we developed a tool which searches for each generated correspondence in the parallel corpus and counts the effective occurrences of such correspondence in the parallel corpus as well the occurrences of the two lexemes of correspondence separately. At the end of the process, we perform a calculation of Log Likelihood ratio to decide to keep or discard the correspondences very rarely used in the corpus.

An other (unexpected) benefit of using a tagged text input

instead of a raw text input, is that we can use the unique identification number that the Fips tagger associates with a lexeme (or with a collocation) in the checking and counting algorithms. We mainly exploited the fact that integer numbers can be used as indexes for arrays speeding up significantly the running of the algorithms (see results in the next section).

4. Results of automatic generation

At this point of the project, we have generated automatically two bilingual lexicons: (1) the English–German lexicon on the basis of English–French and German–French lexicons and (2) the English–Italian lexicon on the basis of English–French and French-Italian lexicons. For the checking of the validity of the correspondences (point 4 of the process) we used the parallel corpus of the debates of the European Parliament during the period 1996 to 2001, EuroParl Version 1 (Koehn, 2005)

The table below summarizes the results. The first four rows correspond to the outputs of the four points of the generation process discussed in section 3. The last one gives the numbers of actual entries:

	English - German	English - Italian
Candidate correspondences	39'806	39'562
Unique correspondences	13'083	17'973
Obtained by collocation pivot	2'233	1'810
Corpus checked correspondences	6'282	7'051
Total validated correspondences	**21'600**	**26'834**

Table 2: Number of entries generated automatically

The quality of derived correspondences is very good but the number of the correspondences that we succeeded to check in the corpus is somehow below our expectation: we found only 26% of the generated correspondences in the EuroParl corpus for English – German and 36% for English - Italian.

The EuroParl corpora we used vary between 17M and 19.6M words depending on the language. The time necessary to tag the corpus by Fips, the time to perform the SQL query that generates the correspondences by transitivity and the time required to check the correspondences in the corpus are given in the tables below. We considered our timing experiments on a 2.67 GHz dual core computer.

English	43.8h
German	37.7h
Italian	26.3h

Table 3: Tagging time

	SQL query	Corpus checking
English – German	0.5 min	3.2 min
English – Italian	0.4 min	3.0 min

Table 4: SQL query and corpus checking time

We may observe that tagging is time-consuming, but this doesn't really matter because tagging is performed only once. What is interesting is that the database query and the checking in the corpus are fast, enabling many runs in order to adjust the parameters of the filtering performed by the query.

5. Conclusion

For convenience (character coding, absence of XML tags, size of corpus) we used EuroParl Version1 corpus. We plan to redo our experiments on the Release 3 of the corpus, comprising 44M words.

The language of the EuroParl corpus is fairly specialized. Many familiar nouns, for instance, are rather unlikely to occur in political debates. We can certainly improve the coverage of the correspondences to check by using also other parallel corpora.

An alternative solution is to check the correspondences against an electronic dictionary.

A third way is to validate the generated correspondences using some semantic information. The WordNet database and the WordNets of other languages will be used for this purpose.

When checking the correspondences in the corpus, we considered a window of one (aligned) line of text in the source and target corpora, i.e. we check if the two lexical items of a correspondence appear in the source, respectively, target aligned line. We want to try a finer window, exploiting the grammatical functions returned by the tagger. For instance, it will allow for a word appearing in the subject position in the source sentence to check if the proposed correspondence is present in the target sentence in the subject position as well.

6. Acknowledgements

7. References

Ahn, Kisuh and Frampton, Matthew. 2006. Automatic Generation of Translation Dictionaries Using Intermediary Languages, in *Cross-Language*

knowledge Induction Workshop of the EACL 06, Trento, Italy, pp 41- 44.

Aymerich, Julia and Camelo, Hermes. 2007. Automatic extraction of entries for a machine translation dictionary using bitexts. in *MT Summit XI*, Copenhagen, pp.21-27 .

Boitet, Christian. 2001. Four technical and organizational keys to handle more languages and improve quality (on demand) in MT, In *MT Summit VIII,* Santiago de Compostela.

Boitet, Christian, Mangeot, Mathieu and Sérasset, Gilles. 2002. The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons, In *Proceedings of the 2nd workshop on NLP and XML, COLING 2002,* Taipei, Taiwan.

Farwell, David, Guthrie, Louise, and Wilks, Yorick. 1992. The Automatic Creation of Lexical Entries for a Multilingual MT System, in *COLING 14: Proceedings of the 14th International Conference on Computational Linguistics*, (Nantes, France), pp. 532-538.

Gamallo Otero, Pablo. 2007. Learning bilingual lexicons from comparable English and Spanish corpora, in *MT Summit XI*, Copenhagen, pp.191-197.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation, In *MT Summit 2005*

Paik, Kyonghee and Shirai, Satoshi and Nakaiwa, Hiromi. 2004. Automatic Construction of a Transfer Dictionary Considering Directionality, In *COLING 2004 Multilingual Linguistic Resources Workshop,* Geneva, pp.25-32

Wehrli, Eric. 2007. Fips, a « deep » Linguistic Multilingual Parser, In *Deep Linguistic Processing Workshop of ACL 2007* , Prague, pp 120-127.

Zhang, Yujie, Ma, Qing and Isahara, Hitoshi. 2007. Building Japanese-Chinese Translation Dictionary Based on EDR Japanese-English Bilingual Dictionary, In *MT Summit XI*, Copenhagen, pp 551-557.