Automatically Generated Customizable Online Dictionaries

Enikő Héja
Dept. of Language Technology
Research Institute for Linguistics, HAS
P.O.Box. 360 H-1394, Budapest
eheja@nytud.hu

Dávid Takács
Dept. of Language Technology
Research Institute for Linguistics, HAS
P.O.Box. 360 H-1394, Budapest
takdavid@nytud.hu

Abstract

The aim of our software presentation is to demonstrate that corpus-driven bilingual dictionaries generated fully by automatic means are suitable for human use. Previous experiments have proven that bilingual lexicons can be created by applying word alignment on parallel corpora. Such an approach, especially the corpus-driven nature of it, yields several advantages over more traditional approaches. Most importantly, automatically attained translation probabilities are able to guarantee that the most frequently used translations come first within an entry. However, the proposed technique have to face some difficulties, as well. In particular, the scarce availability of parallel texts for medium density languages imposes limitations on the size of the resulting dictionary. Our objective is to design and implement a dictionary building workflow and a query system that is apt to exploit the additional benefits of the method and overcome the disadvantages of it.

1 Introduction

The work presented here is part of the pilot project EFNILEX [1] launched in 2008. The project objective was to investigate to what extent LT methods are capable of supporting the creation of bilingual dictionaries. Need for such dictionaries shows up specifically in the case of lesser used languages where it does not pay off for publishers to invest into the production of dictionaries due to the low demand. The targeted size of the dictionaries is between 15,000 and 25,000 entries. Since the completely automatic generation of clean bilingual resources is not possible according to the state of the art, we have decided to provide lexicographers with bilingual resources that can facilitate their work. These kind of lexical resources will be referred to as *proto-dictionaries* henceforward.

After investigating some alternative approaches e.g. hub-and-spoke model (Martin, 2007), alignment of WordNets, we have decided to use word alignment on parallel corpora. Former experiments (Héja, 2010) have proven that word alignment is not only able to help the dictionary creation process itself, but the proposed technique also yields some definite advantages over more traditional approaches. The main motivation behind our choice was that the corpus-driven nature of the method decreases the reliance on human intuition during lexicographic work. Although the careful investigation of large monolingual corpora might have the same effect, being tedious and time-cosuming it is not affordable in the case of lesser used languages.

In spite of the fact that word alignment has been widely used for more than a decade within the NLP community to produce bilingual lexicons e.g. Wu and Xia (1994) and several experts claimed that such resources might also be useful for lexicographic purposes e.g. Bertels et al. (2009), as far as we know, this technique has not been exploited in large-scale lexicographic projects yet e.g. Atkins and Rundell (2008).

Earlier experiments has shown that although word alignment has definite advantages over more traditional approaches, there are also some difficulties that have to be dealt with: The method in itself does not handle multi-word expressions and

[1] EFNILEX is financed by EFNIL

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 51–57,
Avignon, France, April 23 - 27 2012.

the proto-dictionaries comprise incorrect translation candidates, as well. In fact, in a given parallel corpus the number of incorrect translation candidates strongly depends on the size of the proto-dictionary, as there is a trade-off between precision and recall.

Accordingly, our objective is to design and implement a dictionary query system that is apt to exploit the benefits of the method and overcome the disadvantages of it. Hopefully, such a system renders the proto-dictionaries helpful for not only lexicographers, but also for ordinary dictionary users.

In Section 2 the basic generation process is introduced along with the difficulties we have to deal with. The various features of the Dictionary Query System are detailed in Section 3. Finally, a conclusion is given and future work is listed in Section 4.

The proto-dictionaries are available at: `http://efnilex.efnil.org`

2 Generating Proto-Dictionaries – One-Token Translation Pairs

2.1 Input data

Since the amount of available parallel data is crucial for this approach, in the first phase of the project we have experimented with two different language pairs. The Dutch-French language pair represents well-resourced languages while the Hungarian-Lithuaninan language pair represents medium density languages. As for the former, we have exploited the French-Dutch parallel corpus which forms subpart of the Dutch Parallel Corpus (Macken et al., 2007). It consists of 3,606,000 French tokens, 3,215,000 Dutch tokens and 186,945 translation units[2] (TUs). As for Hungarian and Lithuanian we have built a parallel corpus comprising 4,189,000 Hungarian and 3,544,000 Lithuanian tokens and 262,423 TUs. Because our original intention is to compile dictionaries covering every-day language, we have decided to focus on literature while collecting the texts. However, due to the scarce availability of parallel texts we made some concessions that might be questionable from a translation point of view. First, we did not confine ourselves purely to the literary domain: The parallel corpus comprises also philosophical works. Secondly, instead of focusing on direct translations between Lithuanian and Hungarian we have relied mainly on translations from a third language. Thirdly, we have treated every parallel text alike, regardless of the direction of the translation, although the DPC contains that information.

[2]The size of the parallel corpora is given in terms of translation units instead of in terms of sentence pairs, for many-to-many alignment was allowed, too.

2.2 The Generation Process

As already has been mentioned in Section 1, word alignment in itself deals only with one-token units. A detailed description of the generation process of such proto-dictionaries has been given in previous papers, e. g. Héja (2010). In the present paper we confine ourselves to a schematic overview. In the first step the lemmatized versions of each input text have been created by means of morhological analysis and disambiguation[3].

In the second step parallel corpora have been created. We used Hunalign (Varga et al., 2005) for sentence alignment.

In the next step word alignment has been performed with GIZA++ (Och and Ney, 2003). During word alignment GIZA++ builds a dictionary-file that stores translation candidates, i.e. source and target language lemmata along with their translation probabilities. We used this dictionary file as the starting point to create the proto-dictionaries.

In the fourth step the proto-dictionaries have been created. Only the most likely translation candidates were kept on the basis of some suitable heuristics, which has been developed while evaluating the results manually.

Finally, the relevant example sentences were provided in a concordance to give hints on the use of the translation candidates.

2.3 Trade-off between Precision and Recall

At this stage of the workflow some suitable heuristics need to be introduced to find the best translation candidates without the loss of too many correct pairs. Therefore, several evaluations were carried out.

[3]The analysis of the Lithuanian texts was performed by the Lithuanian Centre of Computational Linguistics (Zinkevičius et al., 2005). The Hungarian texts were annotated with the tool-chain of the Research Institute for Linguistics, HAS (Oravecz and Dienes, 2002).

It is important to note that throughout the manual evaluation we have focused on lexicographically useful translation candidates instead of perfect translations. The reason behind this is that translation synonymy is rare in general language e.g. Atkins and Rundell (2008, p. 467), thus other semantic relations, such as hyponymy or hyperonymy, were also considered. Moreover, since the word alignment method does not handle MWEs in itself, partial matching between SL and TL translation candidates occurs frequently. In either case, provided example sentences make possible to find the right translation.

We considered three parameters when searching for the best translations: *translational probability*, *source language lemma frequency* and *target language lemma frequency* (p_{tr}, F_s and F_t, respectively).

The lemma frequency had to be taken into account for at least two reasons. First, a minimal amount of data was necessary for the word alignment algorithm to be able to estimate the translational probability. Secondly, in the case of rarely used TL lemmas the alignment algorithm might assign high translational probabilities to incorrect lemma pairs if the source lemma occurs frequently in the corpus and both members of the lemma pair recurrently show up in aligned units.

Results of the first evaluation showed that translation pairs with relatively low frequency and with a relatively high translational probability yielded cc. 85% lexicographically useful translation pairs. Although the precision was rather convincing, it has also turned out that the size of the resulting proto-dictionaries might be a serious bottleneck of the method (Héja, 2010). Whereas the targeted size of the dictionaries is between 15,000 and 25,000 entries, the proto-dictionaries comprised only 5,521 Hungarian-Lithuanian and 7,007 French-Dutch translation candidates with the predefined parameters. Accordingly, the coverage of the proto-dictionaries should be augmented.

According to our hypothesis in the case of more frequent source lemmata even lower values of translation probability might yield the same result in terms of precision as in the case of lower frequency source lemmata. Hence, different evaluation domains need to be determined as a function of source lemma frequency. That is:

1. The refinement of the parameters yields approximately the same proportion of correct translation candidates as the basic parameter setting,

2. The refinement of the parameters ensures a greater coverage.

Detailed evaluation of the French-Dutch translation candidates confirmed the first part of our hypothesis. We have chosen a parameter setting in accordance with (1) (see Table 1). 6934 French-Dutch translation candidates met the given conditions. 10 % of the relevant pairs was manually evaluated. The results are presented in Table 1. 'OK' denotes the lexicographically useful translation candidates. For instance, the first evaluation range (1st row of Table 1) comprised translation candidates where the source lemma occurs at least 10 times and at most 20 times in the parallel corpus. With these parameters only those pairs were considered where the translation probability was at least 0.4. As the 1st and 2nd rows of Table 1 show, using different p_{tr} values as cut-off parameters give similar results (87%), if the two source lemma frequencies also differ.

F_s	p_{tr}	OK
$10 \leq LF \leq 20$	$p \geq 0.4$	83%
$100 \leq LF \leq 200$	$p \geq 0.06$	87%
$500 \leq LF$	$p \geq 0.02$	87.5%

Table 1: Evaluation results of the refined French-Dutch proto-dictionary.

The manual evaluation of the Hungarian-Lithuanian translation candidates yielded the same result. We have used this proto-dictionary to confirm the 2nd part of our hypothesis, i.e. that the refinement of these parameters may increase the size of the proto-dictionary. Table 2 presents the results. $Expected$ refers to the expected number of correct translation candidates, estimated on the basis of the evaluation sample. 800 translation candidates were evaluated altogether, 200 from each evaluation domain. As Table 2 shows, it is possible to increase the size of the dictionary through refining the parameters: with fine-tuned parameters the estimated number of useful translation candidates was 13,605 instead of 5,521.

F_s	p_{tr}	OK	Expected
$5 \leq$ LF < 30	p > 0.3	64%	4,296
$30 \leq$ LF < 90	p > 0.1	80%	4,144
$90 \leq$ LF < 300	p > 0.07	89%	3,026
$300 \leq$ LF	p > 0.04	79%	2,139
			13,605

Table 2: Evaluation results of the refined Hungarian-Lithuanian proto-dictionary.

However, we should keep in mind when searching for the optimal values for these parameters that while we aim at including as many translation candidates as possible, we also expect the generated resource to be as clean as possible. That is, in the case of proto-dictionaries there is a trade-off between precision and recall: the size of the resulting proto-dictionaries can be increased only at the cost of more incorrect translation candidates.

This leads us to the question of what parameter settings are useful for what usage scenarios? We think that the proto-dictionaries generated by this method with various settings match well different user needs. For instance, when the settings are strict so that the minimal frequencies and probabilities are set high, the dictionary will contain less translation pairs, resulting in high precision and relatively low coverage, with only the most frequently used words and their most frequent translations. Such a dictionary is especially useful for a novice language learner. Professional translators are able to judge whether a translation is correct or not. They might be rather interested in special uses of words, lexicographically useful but not perfect translation candidates, and more subtle cross-language semantic relations, while at the same time, looking at the concordance provided along with the translation pairs, they can easily catch wrong translations which are the side-effect of the method. This kind of work may be supported by a proto-dictionary with increased recall even at the cost of a lower precision.

Thus, the Dictionary Query System described in Section 3 in more detail, should support various user needs.

However, user satisfaction has to be evaluated in order to confirm this hypothesis. It forms part of our future tasks.

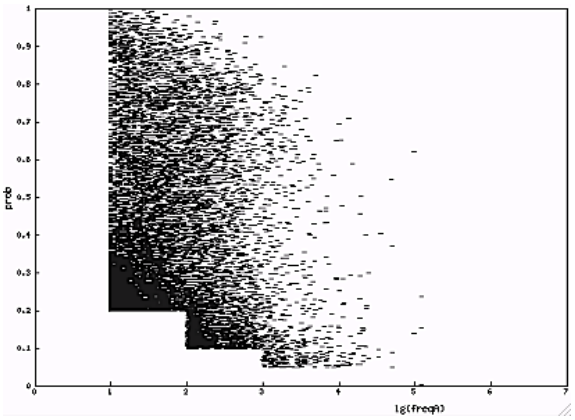

Figure 1: The customized dictionary: the distribution of the Lithuanian-Hungarian translation candidates. Logarithmic frequency of the source words on the x-axis, translation probability on the y-axis.

3 Dictionary Query System

As earlier has been mentioned, the proposed method has several benefits compared to more traditional approaches:

1. A parallel corpus of appropriate size guarantees that the most relevant translations be included in the dictionary.
2. Based on the translational probabilities it is possible to rank translation candidates ensuring that the most likely used translation variants go first within an entry.
3. All the relevant example sentences from the parallel corpora are easily accessible facilitating the selection of the most appropriate translations from possible translation candidates.

Accordingly, the Dictionary Query System presents some novel features. On the one hand, users can select the best proto-dictionary for their purposes on the Cut Board Page. On the other hand, the innovative representation of the generated bilingual information helps to find the best translation for a specific user in the Dictionary Browser Window.

3.1 Customizable proto-dictionaries: the Cut Board Page

The dictionary can be customized on the Cut Board Page. Two different charts are displayed

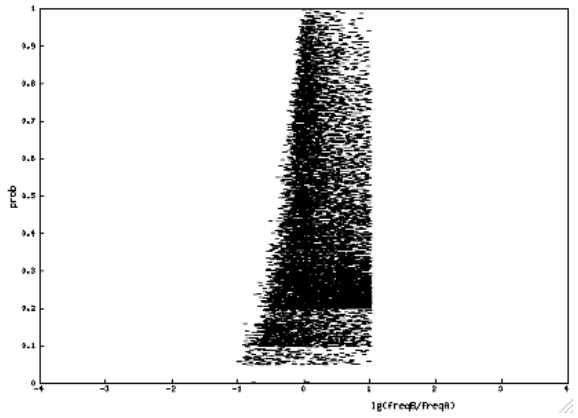

Figure 2: The customized dictionary: the distribution of the candidates. Logarithmic frequency ratio of the source and target words on the x-axis, translation probability on the y-axis.

here showing the distribution of all word pairs of the selected proto-dictionary.

1. Plot 1 visualizes the distribution of the logarithmic frequency of the source words and the relevant translation probability for each word pair, selected by the given custom criteria.

2. Plot 2 visualizes the distribution of the logarithmic frequency ratio of the target and source words and the corresponding translation probability for each word pair, selected by the given custom criteria..

Proto-dictionaries are customizable by the following criteria:

1. Maximum and minimum ratio of the relative frequencies of the source and target words (left and right boundary on Plot 1).

2. Overall minimum frequency of either the source and the target words (left boundary on Plot 2).

3. Overall minimum translation probability (bottom boundary on both plots).

4. Several more cut off intervals can be defined in the space represented by Plot 2: word pairs falling in rectangles given by their left, right and top boundaries are cut off.

After submitting the given parameters the charts are refreshed giving a feedback to the user and the parameters are stored for the session, i. e. the dictionary page shows only word pairs fitting the selected criteria.

3.2 Dictionary Browser

The Dictionary Browser displays four different types of information.

1. List of the translation candidates ranked by their translation probabilities. This guarantees that most often used translations come first in the list (from top to bottom). Absolute corpus frequencies are also displayed.

2. A plot displaying the distribution of the possible translations of the source word according to translation probability and the ratio of corpus ferquency between the source word and the corresponding translation candidate.

3. Word cloud reflecting semantic relations between source and target lemmata. Words in the word cloud vary in two ways.

   First, their *size* depends on their translation probabilities: the higher the probability of the target word, the bigger the font size is.

   Secondly, *colours* are assigned to target words according to their frequency ratios relative to the source word: less frequent target words are cool-coloured (dark blue and light blue) while more frequent target words are warm-coloured (red, orange). Target words with a frequency close to that of the source word get gray colour.

4. Provided example sentences with the source and target words highlighted, displayed by clicking one of the translation candidates.

According to our hypothesis the frequency ratios provide the user with hints about the semantic relations between source and target words which might be particularly important when creating texts in a foreign language. For instance, the Lithuanian lemma *karieta* has four Hungarian eqivalents: ”kocsi” (word with general meaning, e.g. ’car’, ’railway wagon’, ’horse-drown vehicle’), ”hintó” (’carriage’), ”konflis” (’a horse-drawn vehicle for public hire’), ”jármű” (’vehicle’). The various colours of the candidates indicate different semantic relations: the red colour of

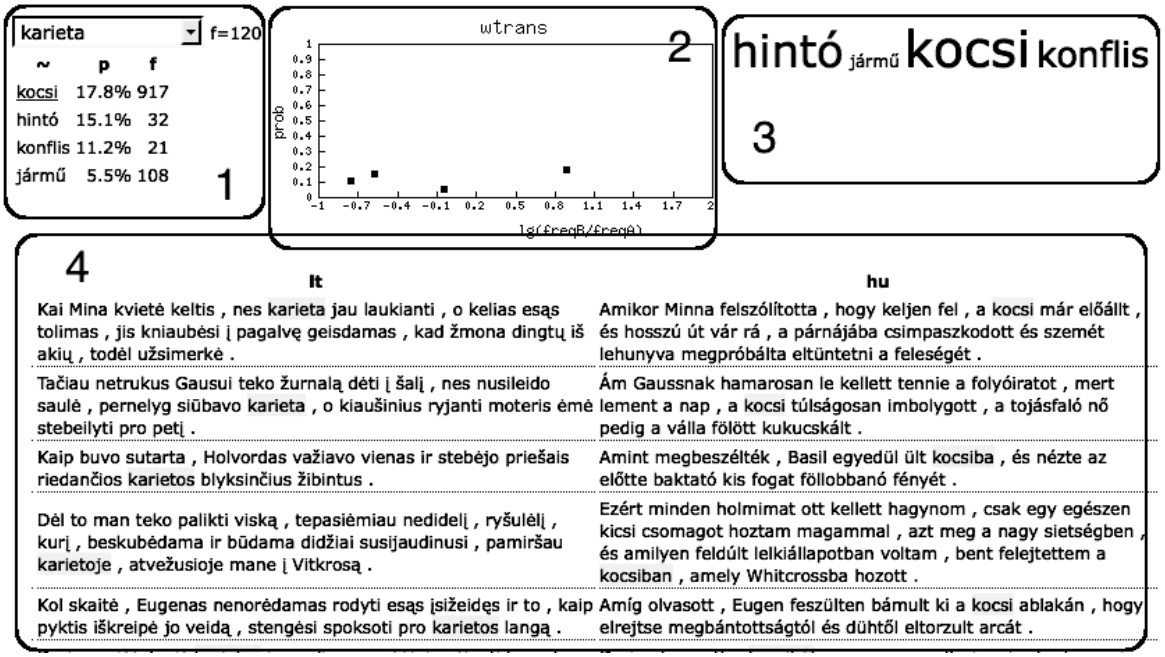

Figure 3: The Dictionary Browser

”kocsi” marks that the meaning of the target word is more general than that of the source word. Conversely, the dark blue colour of ”konflis” shows that the meaning of the target word is more special. However, this hypothesis should be tested in the future which makes part of our future work.

3.3 Implementation

The online research tool is based on the LAMP web architecture. We use a relational database to store all the data: the multilingual corpus text, sentences and their translations, the word forms and lemmata and all the relations between them. The implementation of such a data structure and the formulation of the queries is straightforward and efficient. The data displayed in the dictionary browser as well as the distributional dataset presented on the charts is selected on-the-fly. The size of the database is log-linear with the size of the corpus and the dictionary.

4 Conclusions and Future Work

Previous experiments have proven that corpus-driven bilingual resources generated fully by automatic means are apt to facilitate lexicographic work when compiling bilingual dictionaries.

We think that the proto-dictionaries generated by this technique with various settings match well different user needs, and consequently, beside lexicographers, they might also be useful for end users, both for language learners and for professional translators. A possible future work is to further evaluate the dictionaries in real world use cases.

Some new assumptions can be formulated which connect the statistical properties of the translation pairs, e.g. their frequency ratios and the cross-language semantic relations between them. Based on the generated dictionaries such hypotheses may be further examined in the future.

In order to demonstrate the generated proto-dictionaries, we have designed and implemented an online dictionary query system, which exploits the advantages of the data-driven nature of the applied technique. It provides different visualizations of the possible translations based on their translation probabilities and frequencies, along with their relevant contexts in the corpus. By presetting different selection criteria the contents of the dictionaries are customizable to suit various usage scenarios.

The dictionaries are publicly available at `http://efnilex.efnil.org`.

References

Beryl T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. OUP Oxford.

Ann Bertels, Cédrick Fairon, Jörg Tiedemann, and Serge Verlinde. 2009. Corpus parallèles et corpus ciblés au secours du dictionnaire de traduction. In *Cahiers de lexicologie*, number 94 in Revues, pages 199–219. Classiques Garnier.

Enikő Héja. 2010. The role of parallel corpora in bilingual lexicography. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Lieve Macken, Julia Trushkina, Hans Paulussen, Lidia Rura, Piet Desmet, and Willy Vandeweghe. 2007. Dutch parallel corpus : a multilingual annotated corpus. In *Proceedings of Corpus Linguistics 2007*.

Willy Martin. 2007. Government policy and the planning and production of bilingual dictionaries : The dutch approach as a case in point. *International Journal of Lexicography*, 20(3):221–237.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Csaba Oravecz and Péter Dienes. 2002. Efficient stochastic part-of-speech tagging for hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 710–717, Las Palmas.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Dekai Wu and Xuanyin Xia. 1994. Learning an english-chinese lexicon from a parallel corpus. In *In Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213.

Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 365–370.