

Automatic Methods in the Construction of Bilingual Dictionaries

Enikő Héja

1. Introduction

In what follows I present different approaches to the following questions as part of the EFNILEX project:

- (1) How language technology methods can enable the automatic mapping of dictionary entries between two languages using the existing dictionaries as intermediary resources?
- (2) Research on the automatic construction of bilingual resources from aligned parallel corpora.

Thus, the following report consists of two main parts. The first one considers the first question, based on the work done in the framework of the CLVV project (1993-2007) and described in detail in *International Journal of Lexicography* (vol. 20).

The second part aims at giving an overview of word alignment methods. Word alignment is one general approach which might be used for the generation of bilingual dictionaries.

2. The CLVV project

(1) General Framework of CLVV

In accordance with the objectives of the CLVV project during the period 1993-2007 twenty dictionaries were constructed. In all cases Dutch was either the source or the target language. The average dictionary consists of 45.000 entries with rich microstructures.

From a meta-lexicographic point of view the dictionary projects had two main purposes. Firstly, they aimed at designing dictionaries which are *reversible*, i.e. the source language and the target language are interchangeable. Secondly, they also intended to construct monolingual resources in a way that renders possible the semi-automatic generation of a bilingual dictionary from the monolingual resources. The underlying method is the *hub-and-spoke model* described in detail in subsection (4). Since the hub-and-spoke model presumes the reversibility of bilingual dictionaries, at first I shortly introduce the most important conditions that enabled reversibility.

- (1) The method they used is *linking* instead of *translation* i.e. instead of associating pure word forms (*Form Unit: FU*) they associated form-meaning pairs (*Lexical Unit: LU*), both in the source and the target language.
- (2) The design of the monolingual Dutch database (RBN) is such that it enables the linking method i.e. it consists of LUs.
- (3) OMBI dictionary editor tool facilitates the definition of reversible meaning pairs.

Most of the dictionaries are based on the RBN database (Reference Database of Dutch) and about half of them were built with the OMBI dictionary editor tool. I briefly present these in the next two subsections.

(2) RBN (Referencie Bestand Nederlands 'Reference Database of Dutch')

As described in the paper of van der Vliet (2007), RBN is a Dutch lexical database designed primarily to support the construction of bilingual dictionaries. The monolingual database yields the following benefits: on the one hand, it is reusable in the development of bilingual dictionaries (irrespective of the fact whether Dutch is the source or the target language). On the other hand its structure facilitates the linking method.

RBN consists of a definition and additional properties for each of its 45.000 entries. The properties might be syntactical, morphological, graphemic, semantic (countability, semantic type, systematic polysemy, synonyms), pragmatic and combinatorial (distributional). During the construction of the database great emphasis was laid on the explicit and systematic specification of properties.

The explicit and systematic description of meanings was ensured by the followings:

- (1) They relied on a 38 million-word corpus when differentiating between meanings.
- (2) They aimed at listing only the basic meanings, other senses were derived from the base meanings, for instances by the means of systematic polysemy.
- (3) Nouns were divided into 11 different semantic classes.

(3) OMBI dictionary editor tool

Based on the detailed presentation found in Maks (2007), the main advantage of OMBI dictionary editor tool is that it helps the creation of reversible bilingual dictionaries. This tool was used in about half of the CLVV projects.

The editor comprises three main components:

- (1) Two language components where each language is described as a fully autonomous monolingual resource, without being tailored as a source or a target language. The resources are consisting of form units, lexical units and example units.
- (2) An interlingual component which is a collection of links between lexical units and example units of the two languages.
 - a. Reversibility is guaranteed by the fact that linking operates on lexical units (and not on form units).
 - b. Additional constraints on reversibility might be also defined, therefore unreversible links can be also used, if it is necessary.

Although the editor helps in the development of reversible dictionaries, the need of post-editing still remains. In addition, OMBI does not support either XML or unicode.

(4) Hub-and-spoke Model

The aim of the hub-and-spoke model (Martin 2007) is to generate multilingual dictionaries from reversible monolingual ones. The hub-and-spoke model contains the following steps:

- (1) The first step is the generation of a reversible bilingual dictionary with languages A and B. ($A \leftrightarrow B$)
- (2) The next step is adding a third language in a reversible fashion to the language A, thus generating the links $A \leftrightarrow C$.
- (3) Finally, the links between C and B should be inferred by means of derivation rules, creating the links $B \leftrightarrow C$.

In this case, language A was the *hub language* and languages B and C were the *spoke languages*. Spoke languages therefore are not linked directly to each other but via the hub language.

The theory was put into practice by the derivation of a Danish-Finnish dictionary from the Dutch-Finnish and Dutch-Danish databases made by the means of RBN and OMBI (Laureys 2007). Consequently, in this project Dutch (i.e. RBN) played the role of the hub and Danish and Finish were the spoke languages. The corresponding Danish and Finnish monolingual databases were produced in parallel with the bilingual dictionaries. The most important requirements for a successful merging were met:

- (1) The entries of spoke language databases are of the same structure.
- (2) They showed the same level of semantic specification.
- (3) The core of the example units is largely alike.

The linking process between the spoke languages is semi-automatic because post-editing is needed, but the amount of labour is reduced drastically.

(5) Conclusions

The main advantage of reversibility is that it facilitates the construction of bilingual dictionaries. The same might be said of resources constructed to meet the requirements of the hub-and-spoke model. However, post-editing is needed when applying these policies to the task of dictionary construction.

From our point of view the most important thing is that the hub-and-spoke model requires databases which meet certain expectations. The development of such databases is a time-consuming task.

3. Extraction of Bilingual Lexicons and Word Alignment

As we saw above, the hub-and-spoke model presumes the existence of monolingual databases with approximately the same structure and size. There are also certain statistical methods that can be used for the purpose of dictionary extraction without such refined resources. These methods seek a solution for the problem of word alignment. Word alignment aims at finding alignment links between words in a parallel corpus. Bilingual lexicon extraction goes further: its goal is to identify the lexical word type links based on alignment between word tokens. Thus, dictionary extraction might be decomposed into two basic steps:

- (1) The text/sentence alignment of the parallel corpus is extended to a word alignment.
- (2) Some criterion is used (e. g. frequency) to select the aligned pairs for which there is enough evidence to include them in a bilingual dictionary.

Usually bilingual lexicon extraction is conceived as an easier task than word alignment, since uncertain relations and translation irregularities could be excluded from the extracted lexicon.

That is why we need to differentiate between the evaluation results of a word alignment system and a dictionary extraction system. However, since in the literature most results concern the evaluation of word alignment systems (Rada et al. 2003, Martin et al. 2005), and there is a correlation between the performance of a word alignment algorithm and the resulting dictionary in this paper we are going to pay more attention to the performance of word alignment systems.

4. Word Alignment and the Task

Since there is no decision about the specific languages to be included in the dictionary, my intention was to give a general picture of word alignment. Therefore, I confine myself to the presentation of language-independent methods. These rely on purely statistical information, and do not make use of previous language-specific knowledge (i.e. POS information, phrasal information or syntactic dependence, etc.). On the other hand, since concrete results are dependent on the size and type of input data, instead of emphasizing the concrete evaluation results of various word alignment algorithms and systems, I would like to give a more detailed description of advantages and drawbacks of different algorithms.

First of all, some introductory remarks are in order concerning some general questions or pre-requisites of word alignment systems.

- (1) What kind of resources are at disposal for the language pair to be aligned? Most algorithms presume sentence-aligned parallel corpora. If there is no parallel corpora for the given language pair:
 - a. Parallel corpora might be produced. As Tiedemann (2003) points out, there are reliable algorithms for sentence alignment.
 - b. Some word alignment approaches does not rely on sentence-alignment (e.g. K-vec++ algorithm by Pedersen and Varma, 2002)
- (2) Is it sufficient to achieve only high precision even on the coast of low recall, or recall should be taken also into consideration? (Some simple methods in principle are not able to achieve 100% recall. The competitive linking algorithm (Melamed, 1997) is an example.)
- (3) All approaches require tokenization.
- (4) The morpho-syntactic properties of a language could be also important.
 - a. In the case of highly inflective languages morphological preprocessing is needed to avoid the problem of data sparseness.
 - b. Most algorithms rely on a version of the one-to-one assumption. If the strict version of the assumption holds, without morphological preprocessing it is impossible to recognize word pairs such as *aller* and *to go*. For the same reason previous recognition of multi-word-expressions (such as collocations, idiomatic expressions, phrases) can be also substantial.

- (5) Are there already existing dictionaries for the given language pair? Already existing dictionaries might improve the performance of the algorithm.

5. Algorithms of Word Alignment

Word alignment methods enable the unsupervised learning of word pairs from sentence-aligned corpora. These algorithms can be classified into two broad categories: *association approaches* and *estimation approaches*. In what follows, I will give a short overview of the basic properties of both kinds of word alignments.

(1) Association Approaches

Association approaches are also called *heuristic approaches* or *hypothesis testing approaches* in the literature. As noted by Och and Ney (2003) a common idea behind statistical association measures is to test if two words co-occur significantly more often than it would be expected if they would co-occur purely by chance. To test the independence hypothesis they count co-occurrence frequencies in the aligned regions and use some association measure to determine how independent two words are. Most frequently the Dice coefficient or some variants of it are used. Varma (2002) gives a detailed comparison of various measures and tests of association such as Pointwise Mutual Information, Dice coefficient, Log-likelihood Ratio, Pearson's Chi-square test, Odds ratio, T-score and Fischer's Exact Test. Ribeiro, Lopez and Mexia (2000) consider the performance of 23 similarity measures in a dictionary extraction context.

In the next step the best translation candidate is selected based on some suitable heuristics.

Manning and Schütze (1999), among others, emphasize that association approaches can be misled in situations where a word in the source language frequently occurs with more than one word in the target language. For instance, based on the Hansard corpus *house* is wrongly associated with *communes* instead of *chambre*, since in this corpus *house* most frequently occurs with the expression *chambre de communes*. An alignment between *house* and *communes* is called *indirect association*.

One solution to this problem is Melamed's (2000, 1997) *competitive linking algorithm*, which at first finds the highest ranking association score (i, j) in the association matrix, align these two word positions and removes the corresponding row and column from the matrix. This procedure is iterated until every source or target language word is aligned. The result of this approach is that indirect associations occur less often. Therefore, this approach is able to give higher precision results than other association methods, but only on the cost of a relatively low recall. The one-to-one restrictive nature of this method excludes many translation phenomena. In this case linguistic preprocessing (recognition of multi-word expressions) is able to compensate the negative correlates of the method.

(2) Estimation approaches for word alignment

Estimation approaches of word alignment are inspired by statistical machine translation. Statistical machine translation is an application of the noisy channel model from information theory (Shannon 1948) to the task of machine translation. In what follows, I will

give a brief outline of how the noisy channel model can be used for the purpose of word alignment based on Tiedemann (2003), Hiemstra (1996) and Manning and Schütze (1999).

When the noisy channel model is applied to the task of machine translation, the source language S and the target language T are considered to be random variables that produce sentences. Translation is modeled as a transmission of a source language sentence \mathbf{s} through a noisy channel that transforms it into a \mathbf{t} target language sentence. In a noisy channel model the target sentence \mathbf{t} is considered to be the observable part of the system and the task of the model is to find the original input string \mathbf{s} that has been transmitted through the channel in order to produce \mathbf{t} target sentence. Therefore, our goal is to determine the most probable source language sentence, $\hat{\mathbf{s}}$, given \mathbf{t} target language sentence:

$$(E1) \quad \hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} P(\mathbf{s}|\mathbf{t}) = \underset{\mathbf{s}}{\operatorname{argmax}} \frac{P(\mathbf{t}|\mathbf{s})P(\mathbf{s})}{P(\mathbf{t})}$$

Because $P(\mathbf{t})$ is independent of \mathbf{s} and, therefore, is constant for all possible source language sentences, $P(\mathbf{t})$ might be omitted from the equation. So, the basic equation is as follows:

$$(E2) \quad \hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} P(\mathbf{t}|\mathbf{s})P(\mathbf{s})$$

Here the probability $P(\mathbf{s})$ is the *prior probability* or the *language model*. It expresses the probability that the translator will translate the source language sentence \mathbf{s} .

The distribution $P(\mathbf{t}|\mathbf{s})$ – the *translation model* – can be looked upon as a source language – target language dictionary (one that has information about all possible sentences in the source and the target language), thus it gives a relatively high probability to sentences that are each others translation.

Both distributions $P(\mathbf{s})$ and $P(\mathbf{t}|\mathbf{s})$ are enormously complex. The next step is to define models for these probability distributions.

One important property of estimation approaches is how $P(\mathbf{s})$ is modeled. A possible solution is to presume that words are independent in the sentences and estimate $P(\mathbf{s})$ as the product of the probability of each word in the sentence. That is, assuming that \mathbf{s} sentence consists of l words, s_1, s_2, \dots, s_l , $P(\mathbf{s}) = P(s_1) \cdot P(s_2) \cdot \dots \cdot P(s_l)$ might be a good estimation. The parameters of these models are then estimated based on the corpora. In this case the model ignores any sequence and position information, thus, this is a zero-order model. As Osch and Ney (2003) showed higher-order models yield better results, which might be a consequence of the fact that position information is inherently present in all natural languages.

Suppose that the source language sentence \mathbf{s} is defined as a sequence of l words: $s_1 s_2 \dots s_l$ and the corresponding target language sentence \mathbf{t} is $t_1 t_2 \dots t_l$. In this case the translation model $P(\mathbf{t}|\mathbf{s})$ might be estimated by the following equation:

$$(E3) \quad P(\mathbf{t}|\mathbf{s}) = P(t_1|s_1) \cdot P(t_2|s_2) \cdot \dots \cdot P(t_l|s_l)$$

(E3) estimation presumes that one target language word is generated by one source language word, which is the one-to-one assumption found also in Melamed's competitive linking algorithm. However, there are some models which are able to deal with the phenomenon when more than one target language word is generated by a single source word. The notion of *fertility* is introduced to handle such cases.

The basic idea behind the translation model is to conceive word alignment as a hidden variable. The task now is the estimation of translation probabilities $P(t_i|s_k)$ based on data only with sentence alignment. Several methods utilize the *expectation maximization* algorithm for the maximum likelihood estimation of translation probabilities based on incomplete data.

The EM algorithm starts with the random initialization of translation probabilities and updates them iteratively by maximizing the likelihood function until the process converges at a maximum. The correctness of the algorithm is proved in Dempster et al (1977).

The EM algorithm was firstly introduced to analyze parallel corpora by Brown et al (1990).

Most estimation approaches are the extensions of the IBM models described in Brown et al. (1993) such as the algorithm described in Hiemstra (1996) or Model 6 described in Och and Ney (2003). Och and Ney (2003) gives also a detailed comparison of the different IBM models and concludes that alignment models with a first-order dependence and a fertility model yields significantly better results than simple heuristic models and estimation approaches with zero-order dependence. There is also a free software GIZA++ (designed and written by Och) for training purposes.

6. Advantages and disadvantages

According to Och and Ney (2003) the main advantage of heuristic models is their simplicity. They are easy to implement and understand. But the specific similarity function seems to be completely arbitrary. They showed in their paper that certain kinds of statistical models yield significantly better results than simple heuristic methods.

Accordingly, statistical alignment models are more coherent. The general principle for coming up with an association score between words results from a statistical estimation theory and the parameters of the models are adjusted such that the likelihood of the models on the training corpus is maximized. Tiedemann (2003) notes that different alignment strategies might be chosen to suit particular needs. Competitive linking algorithm is appropriate for the task of lexicon extraction as it yields high precision results among one-to-one links, while estimation methods should be preferred when coverage also plays a great role, such as machine translation.

7. The Use of External Resources

Although Manning and Schütze (1999) held high expectations for machine-readable bilingual dictionaries in word alignment, according to Tiedemann (2003) the impact of such resources depends very much on their size and appropriateness with respect to the corpus and its domain. In their experiment Och and Ney found that the improvement¹ contributed by the use of a bilingual dictionary negatively correlates with the size of the training corpus and is small compared to the improvement achieved through the use of better alignment models.

Language-specific expert knowledge might be also used as an external resource for word alignment. I briefly introduced some of these in section 4.

¹ In terms of *alignment error rate*. For further details see the next section.

8. Evaluation

Tufis et al. (2004) consider the task of word alignment more difficult than that of dictionary extraction, since in the latter one translation pair is considered correct if there is at least one context in which it has been correctly observed. Thus, multiply-occurring pairs would count only once for the final lexicon. While in word alignment each occurrence of the same pair counts equally. As mentioned in the introduction, since the results of a lexicon extraction task depend on the results of the corresponding word alignment (and vice versa), moreover, there are much more evaluation results of word alignment tasks, in this paper we confine ourselves to the evaluation of word alignment.

First of all, I briefly present an evaluation measure, *alignment error rate* (AER), introduced by Och and Ney (2003) and the corresponding evaluation scheme.

Annotators are asked to distinguish between two kinds of annotations: *sure alignments* (S), when the alignments are unambiguous and *possible alignments* (P) for ambiguous alignments (idiomatic expressions, free translations and missing function words). The quality of an alignment A is computed then by the redefined precision and recall measures.

$$(E4,E5) \quad \text{recall} = \frac{|A \cap S|}{|S|}, \text{ precision} = \frac{|A \cap P|}{|A|}$$

Accordingly, a recall error can occur only if an S alignment is not found and a precision error can occur only if the found alignment is not even P. The extended version of F-measure is called *alignment error rate*:

$$(E6) \quad \text{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

Model	Training scheme	Size of training corpus			
		0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1 ^S	40.6	33.6	28.6	25.9
Model 2	1 ^S 2 ^S	46.7	29.3	22.0	19.5
HMM	1 ^S H ^P	26.3	23.3	15.0	10.8
Model 3	1 ^S 2 ^S 3 ^S	43.6	27.5	20.5	18.0
	1 ^S H ^P 3 ^S	27.5	22.5	16.6	13.2
Model 4	1 ^S 2 ^S 3 ^S 4 ^S	41.7	25.1	17.3	14.1
	1 ^S H ^P 3 ^S 4 ^S	26.1	20.2	13.1	9.4
	1 ^S H ^P 4 ^S	26.3	21.8	13.3	9.3
Model 5	1 ^S H ^P 4 ^S 5 ^S	26.5	21.5	13.7	9.6
	1 ^S H ^P 3 ^S 4 ^S 5 ^S	26.5	20.4	13.4	9.4
Model 6	1 ^S H ^P 4 ^S 6 ^S	26.0	21.6	12.8	8.8
	1 ^S H ^P 3 ^S 4 ^S 6 ^S	25.9	20.3	12.5	8.7

The table from Och and Ney (2003) nicely presents that the performance of different systems (in terms of AER) depends primarily on the refinement of the model and the size of the training corpus. Accordingly, Model 6 trained on the largest corpus yielded the lowest AER score, while the simple Dice-coefficient with the smallest corpus performed the worst.

In the Romanian-English word align task (Martin et al 2005) a system based on an IBM model gave the lowest alignment error rate and the highest precision was achieved by a system based on similar grounds.

References:

- Brown, P., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Rossin, P.: A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85., 1990.
- Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L.: The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313., 1993.
- Dempster, A. P., N. M. Laird, and D. B. Rubin.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977. 39(1):1–22.
- Hiemstra, D.: Using statistical methods to create a bilingual dictionary', *Master's Thesis, University of Twente*, 1996.
- Laureys, G.: Optimizing Procedures for the Making of Bilingual Dictionaries and the Concept of Linking Contrastive Lexical Databases, *International Journal of Lexicography*, September 1, 2007; 20(3): 295-311.
- Maks, I.: Ombi: The Practice of Reversing Dictionaries, *International Journal of Lexicography*, September 1, 2007; 20(3): 259-274.
- Manning D. C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press, Massachusetts, US, 1999.
- Martin, J., Mihalcea, R., Pedersen, T.: Word Alignment for Languages with Scarce Resources, in: *Proceedings of the ACL Workshop on Building and Exploiting Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, MI, June 2005.
- Martin, W.: Government Policy and the Planning and Production of Bilingual Dictionaries: The 'Dutch' Approach as a Case in Point, *International Journal of Lexicography*, September 1, 2007; 20(3): 221 - 237.
- Melamed, D.: Models of Translational Equivalence among Words, *Computational Linguistics*, Vol. 26, No. 2. 221-249., 2000.
- Mihalcea, R. and Pedersen, T.: An Evaluation Exercise for Word Alignment, in *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May 2003.
- Och, F. J., and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.
- Ribeiro, A., Pereira Lopes, J., G., Mexia, J.: Extracting Equivalents from Aligned Parallel Texts: Comparison of Measures of Similarity. *IBERAMIA-SBIA*: 339-349, 2000.
- Shannon, C. E.: A mathematical theory of communication, *Bell Systems Technical Journal* 27, 1948, 379-423.
- Tiedemann, J.: Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Doctoral Thesis, *Studia Linguistica Upsaliensia I*, ISSN 1652-1366, ISBN 91-554-5815-7
- Tufiş, D., Barbu, A. M., Radu I.: Extracting Multilingual Lexicons from Parallel Corpora, *Computers and the Humanities*, Volume 38, Issue 2, May 2004, Pages 163 - 189, ISSB 0010-4817.
- Varma, N.: Identifying Word Translations in Parallel Corpora Using Measures of Association, *Master's Thesis, University of Minnesota*, 2002.
- Vliet, van der, H.: The Referentiebestand Nederlands as a Multi-Purpose Lexical Database *International Journal of Lexicography*, September 1, 2007; 20(3): 221-238