Deriving a Bilingual Lexicon for Cross Language Information Retrieval

D. Hiemstra

Department of Computer Science, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
E-mail: hiemstra@cs.utwente.nl

Abstract

In this paper we describe a systematic approach to derive a bilingual lexicon automatically from parallel corpora. Following this approach, a lexicon was derived from the English and Dutch version of the Agenda 21 corpus. With the lexicon and a part of the corpus that was not used to derive the lexicon, a bilingual retrieval environment was build. Recall and precision of monolingual (Dutch) retrieval was compared to recall and precision of bilingual (Dutch-to-English) retrieval. An experiment was conducted with the help of eight naive users who formulated queries and judged the relevance of retrieved fragments. The experiment shows 78% precision and 51% relative recall of monolingual retrieval, against 67% precision and 82% relative recall of bilingual retrieval.

Keywords: Lexical Acquisition, Parallel Corpora, Statistical Natural Language Processing, Cross Language Information Retrieval.

1 Introduction

The recent enormous increase in the use of information from internet and cd-rom has led to databases being available in many languages. Often the relevance of the documents in these databases goes beyond the scope of a region or country. In cases where the documents are only available in a foreign language, multilingual document retrieval provides access for people who are non-native speakers of the foreign language or not a speaker of the language at all. A machine translation system that is capable of translating document indexes and queries can be used to identify documents for human translation.

In this paper we describe a systematic approach to derive a *probabilistic* bilingual lexicon automatically from a parallel corpus. A probabilistic bilingual lexicon assigns a probability value to each possible translation of an entry (see table 1). The bilingual lexicon can be used directly as a statistical translation tool or to enhance an existing general purpose machine translation system with domain dependent translations.

sustainable	
duurzame	0.80
duurzaam	0.20

Table 1: an example entry

Recently much research was done into aligning bilingual corpora at the sentence level [1, 2, 3]. We used a program published by Gale and Church [2] to align the sentences of a bilingual corpus. The program makes use of the fact that longer sentences tend to be translated into longer sentences, and shorter sentences tend to be translated into shorter ones. Throughout the rest of this paper we will use the fact that we know the translation of each sentence in the corpus, but not the translation of the words.

To align the words of each sentence also, a statistical algorithm called the Expectation Maximisation algorithm (EM-algorithm) was used. The EM-algorithm was proofed to be correct by Dempster, Laird and Rubin in 1977 [4] and was first used to analyse bilingual corpora at IBM in 1990 [5, 6]. The IBM article inspired many research centres over the world to use statistical methods for machine translation also. This paper contributes to this discussion in two new ways:

(*i*) An EM-algorithm was developed that compiles a bi-directional lexicon (that is, a lexicon that can be used to translate from for example English to Dutch as well as from Dutch to English). We believe that there are two good reasons to conduct a bi-directional approach. Firstly, a bi-directional lexicon will need less space than two uni-directional lexicons. Secondly, we believe that a bi-directional approach will lead to better estimates of the translation probabilities than the uni-directional approach.

(*ii*) We built a document retrieval environment and compared recall and precision of a monolingual (Dutch) retrieval engine to recall and precision of a bilingual (Dutch-to-English) retrieval engine. The bilingual lexicon, compiled with the EM-algorithm from a parallel corpus, was used to translate Dutch queries automatically to corresponding English queries. The experiment was conducted with the help of eight naive users who formulated the queries and judged the relevance of the retrieved documents.

The rest of the paper is organised as follows: In section 2 we will give an informal description of the probability model and the estimating algorithm. Section 3 gives a detailed description of the conducted experiments. Finally, section 4 will conclude with some final remarks.

2 Assigning probabilities to translations

The problem of modelling the translation of sentences may, to some extend, be compared to problems in medicine and social sciences. In much of these studies a population of people is categorised in for example, whether a smoker or not and different types of cancer. Frequently the physician collecting such data is interested in the relationships or associations between pairs of such categorical data.

We will do something like that in this paper. Suppose we want to study the bilingual corpus of table 2 that consists of four pairs of Dutch and four pairs of English sentences which are each others translation.

He waits.	*Hij wacht.*
you wait.	*jij wacht.*
he can.	*hij kan.*
you can.	*jij kunt.*

Table 2: an example corpus

A statistical approach will be taken. We assume that the corpus consists of randomly drawn samples of English-Dutch translations. To every pair of sentences (E, D) in the corpus a probability measure $P(E, D)$ will be assigned to be interpreted as the probability that the sentence pair (E, D) appears in our corpus as a translation pair. We expect the probability $P(E, D)$ to be very small for pairs like (*hij wacht, you can*) and relatively large for pairs like (*jij kunt, you can*) In the following sections three aspects of the approach will be emphasized. Section 2.1 defines how to model the observations of sentence pairs in the corpus. In section 2.2 the EM-algorithm will be defined to estimate the unknown parameters of the model. Section 2.3 gives a method to approximate the E-step of the algorithm in reasonable time.

2.1 The translations model

Just like the physician has to diagnose the condition of the patient he examines ("what type of cancer does this patient have?"), we will assign an equivalence class to every word we observe. If some form of morphological analysis is performed we might want to assign the words *wait* and *waits* to the same equivalence class. Between words that fall into the same equivalence class exists an equivalence relation, i.e. the words share a certain property. In table 2 the words *wait* and *waits* share the same meaning.

In the experiment, as little linguistic knowledge as possible was used. Therefore every different word was assigned to a separate equivalence class, so for example morphological related words like *wait* and *waits* were treated as two (entirely) different words. Case-distinction of letters will not be made. For example, the words *he* and *He* were assigned to the same equivalence class. In the example corpus, there are five different English words and also five different Dutch words. This makes a total of twenty-five possible translations. that will be displayed in a so-called contingency table.

	he	*waits*	*you*	*wait*	*can*	
hij	n_{11}	n_{12}	$\cdots$		n_{1c}	$n_{1.}$
wacht	n_{21}				$:$	$n_{2.}$
jij	$:$					$:$
kan						
kunt	n_{r1}	$\cdots$			n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	$\cdots$		$n_{.c}$	$n_{..}$

Table 3: contingency table for the example corpus

The cell frequencies n_{ij} in the table represent the number of times the English word i and the Dutch word j are each others translation in the corpus. The marginal totals $n_{i.}$ represent the number of times the English word i appears in the corpus. The marginal totals $n_{.j}$ represent the number of times the Dutch word j appears in the corpus. In terms of cell frequencies n_{ij} the marginal totals are given by:

$$n_{i.} = \sum_{j=1}^{5} n_{ij}, \;\; n_{.j} = \sum_{i=1}^{5} n_{ij} \qquad (1)$$

Each cell frequency n_{ij} will be assigned an unknown probability parameter p_{ij} which is the probability that the English word i and the Dutch word j appear in the corpus as a translation pair. To define the probability measure P as a function of the observations n_{ij} and the parameters p_{ij} it is assumed that the word translation pairs in a sentence pair 'appear' independently of each other. Furthermore a sentence is modelled as a collection of words, i.e. there is no sequence between words or translation pairs of words. Finally we assume that each word in one language is alligned to only one word in the other language, and vice versa. These three assumptions lead to the following model of the probability measure P:

$$P(N = n_{11} \cdots n_{rc}) = \frac{n_{..}!}{n_{11}! \cdots n_{rc}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}} \quad (2)$$

Equation 2 is the well known multinomial distribution and its unknown parameters p_{ij} form the probabilistic bilingual lexicon we are looking for. The estimate $\hat{p}_{ij}$ of p_{ij} that makes the observations as likely as possible is given by

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{..}} \quad (3)$$

which is the maximum likelihood estimate of the unknown parameters.

2.2 The EM-algorithm

Every observation in the parallel corpus must be represented by Table 3. However, the information needed to fill table 3 is not explicitly present in the observations. The observations are *incomplete*, i.e. the marginal totals $n_{i.}$ and $n_{.j}$ of the cell frequencies n_{ij} are known, but the cell frequencies themselves are unknown. Table 4 shows the incomplete observation of the first sentence in the example corpus. For convenience, cell frequencies that are 0 are not displayed.

	he	*waits*	*you*	*wait*	*can*	
hij	?	?	-	-	-	1
wacht	?	?	-	-	-	1
jij	-	-	-	-	-	0
kan	-	-	-	-	-	0
kunt	-	-	-	-	-	0
	1	1	0	0	0	2

Table 4: incomplete observation of (*he waits*, *hij wacht*)

Given the marginal totals $n_{i.}$ and $n_{.j}$ and the probability parameters p_{ij} it is possible to compute the expected values of the cell frequencies $E(N|n_{1.} \cdots n_{r.}, n_{.1} \cdots n_{.j}, p_{11} \cdots p_{rc})$. This seems a serious problem. Without the probability parameters p_{ij}, the expected cell frequencies n_{ij} cannot be calculated. Without the cell frequencies n_{ij}, the maximum likelihood estimate of the probability parameters p_{ij} (equation 3) cannot be calculated.

From the definition of the EM-algorithm [4] the following iterative solution can be constructed.

(i) Take an initial estimate of the probability parameters.

(ii) *Expectation-step*: For each sentence, calculate the expected cell frequencies given the marginal totals and the probability parameters.

(iii) *Maximisation-step*: Add the expected observations and calculate the maximum likelihood estimate as defined by equation 3.

(iv) Repeat (ii) and (iii) until the probability parameters do not change significantly anymore.

If no linguistic knowledge is used, initially every word pair is equally likely as a translation. For the example corpus of table 2 the initial estimate then must be $p_{ij} = \frac{1}{25}$ for each possible i and j. Table 5 and 6 give an impression of the way the algorithm behaves on the simple example corpus of table 2. After five iterations of the algorithm the parameters of the model do not change significantly anymore.

	he	*waits*	*you*	*wait*	*can*	
hij	1.0	0.5	-	-	0.5	2
wacht	0.5	0.5	0.5	0.5	-	2
jij	-	-	1.0	0.5	0.5	2
kan	0.5	-	-	-	0.5	1
kunt		-	0.5	-	0.5	1
	2	1	2	1	2	8

Table 5: expected complete observation of the corpus in the first iteration

	he	*waits*	*you*	*wait*	*can*	
hij	2.0	-	-	-	-	2
wacht	-	1.0	-	1.0	-	2
jij	-	-	2.0	-	-	2
kan	-	-	-	-	1.0	1
kunt		-	-	-	1.0	1
	2	1	2	1	2	8

Table 6: expected complete observation of the corpus in the fifth iteration

2.3 Approximation of the E-step

The EM-algorithm we defined in the previous section works fine on the example corpus of table 2 because it has very short sentences. Calculating the expectation of a random variable requires summing over all possible values of that random variable. The number of possible alignments of a sentence pair increases exponentially with the length l of both sentences. The average number of words per sentence in the corpus that was used in the experiments is more than 20. On a sentence pair of which both sentences have the same length l, the number of possible alignments is $20! > 10^{18}$. Because this cannot be calculated in reasonable time, it is necessary to approximate the E-step of the algorithm.

A very old way to find missing values in $l \times l$ contingency table is the *iterative proportional fitting procedure* (IPFP). This algorithm was probably first described in 1937 by Kruithof [7]. The basic IPFP takes a contingency table with initial cell frequencies $n_{ij}^{(0)}$ and sequentially scales the table to satisfy the observed data $m_{i.}$ and $m_{.j}$. We assume that the marginal totals $n_{i.}^{(0)}$ and $n_{.j}^{(0)}$ are not yet in correspondence with the observed data $m_{i.}$ and $m_{.j}$. The pth iteration of the algorithm consists of two steps which form:

$$\begin{aligned} n_{ij}^{(p,1)} &= n_{ij}^{(p-1,2)} \cdot m_{i.} / n_{i.}^{(p-1,2)} \\ n_{ij}^{(p,2)} &= n_{ij}^{(p,1)} \cdot m_{.j} / n_{.j}^{(p,1)} \end{aligned} \qquad (4)$$

The first superscript refers to the iteration number, and the second to the step number within iterations. The algorithm continues until the observed data $m_{i.}$ and $m_{.j}$ and the marginal totals $n_{i.}^{(p)}$ and $n_{.j}^{(p)}$ are sufficiently close. The IPFP will work on a partial contingency table containing only those words that appeared in the sentence pair.

Succes of the approximation depends heavily on the initial estimate of the IPFP. An initial estimate that proofed to be succesfull is based on taking together equivalence classes.

$$n_{ij}^{(0)} = \frac{p_{ij}(p_{..} - p_{i.} - p_{.j} + p_{ij})}{(p_{i.} - p_{ij})(p_{.j} - p_{ij})} \qquad (5)$$

The marginal probability parameters $p_{i.}$ and $p_{.j}$ are defined according to the marginal totals in equation 1.

2.4 Discussion

The algorithm we constructed combines EM with IPFP. Both algorithms are commonly used for maximum likelihood estimation in log-linear models. In this paper IPFP is used to replace the exact calculation of the E-step of EM by an approximation that satisfies the observed marginal totals. We do not have theoretical proof that, given the observations in the corpus, the algorithm indeed converges to the most likely solution of the model. However, preliminary results [8] indicate better performance than the algorithm developed at IBM.

3 Experimental results

To test the performance of the algorithm, we compiled a bilingual probabilistic lexicon from the parallel English and Dutch version of *Agenda 21.* Agenda 21 is an international document available in more than 80 languages reflecting the results of the United Nations Conference on ecology in 1992 in Rio de Janeiro. It contains guiding principles for sustainable development, covering topics such as deforestation, biological diversity, etc.

Only half the corpus was used to derive the lexicon. To test the lexicon in a bilingual retrieval environment, the other half of Agenda 21 was used to build a bilingual document base. Recall and precision of a monolingual (Dutch) retrieval engine were compared to recall and precision of a bilingual (Dutch-to-English) retrieval engine. The following sections give a detailed description of the conducted experiments.

3.1 Compiling the bilingual lexicon

The training corpus consisted of 4664 parallel sentences. With the training corpus, a bilingual lexicon was compiled consisting of 3854 English words and 5462 Dutch words. More than 21 million unknown parameters were estimated.

The algorithm presented above expects the English and the Dutch sentences to be of equal length. Because the sentences almost never have the same length, the assumption was made that some words are not translated at all. To model this assumption, a special (*null*) word was introduced for each language. If the length of, for example, the English sentence was smaller than the length of the parallel Dutch sentence, the English sentence was filled up with the special (*null*) words.

the		*de*	
de	0.68	*the*	0.49
het	0.14	(*null*)	0.38
(*null*)	0.03	*of*	0.02
in	0.01	*to*	0.01
te	0.01	*as*	0.01
aanmerking	0.01	*in*	0.01

Table 7: example entries of function words

Table 7 gives an examples of the results of the algorithm on an English and a Dutch function word after six training steps. The six most probable translations of the entry are displayed, together with the probability of each possible translation. The algorithm found out that the definite article *the* has the two most probable Dutch translations *de* and *het*, which is, by the way, correct. It is also quite probable that a definite article in either language is not translated in the other language. Still, some errors are made with low probability. Table 8 shows how the algorithm handles morphologically related words and synonyms. Morphology and synonyms often get special attention in information retrieval systems.

local		*duurzame*	
plaatselijke	0.51	*sustainable*	0.93
lokale	0.24	*unsustainable*	0.02
lokaal	0.15	*renewable*	0.02
plaatselijk	0.09	*consumption*	0.01
maken	0.01	*sustainability*	0.01

Table 8: example entries of morphologically related words and synonyms

Finally, table 9 shows example entries of translations that cannot be modeled very well by the approach taken in this paper. In Dutch, nouns can be compounded to form new words. For example the Dutch word *volksgezondheid* is a compound and should be translated as *people's health.* Because nouns are usually not compounded in English, the algorithm will find only a partially correct translation.

volksgezondheid		*health*	
health	1.00	*gezondheid*	0.28
		gezondheidszorg	0.20
		volksgezondheid	0.11
		gezondheidsprobl.	0.05
		gezondheids	0.04
		te	0.02

Table 9: example entries of compound nouns

3.2 Cross language retrieval results

The experiment was conducted with the help of eight naive users that retrieved fragments from the bilingual document base. The bilingual document base contained 1545 English fragments of Agenda 21 that were not used to compile the lexicon, together with their parallel Dutch translations. The native language of the volunteers was Dutch. The experiment was organised as follows:

(*i*) Each volunteer was asked to formulate five Dutch queries on the domain of sustainable development and ecology.

(*ii*) Each query was used to retrieve documents from the Dutch database, using a Boolean retrieval model and word-based indexing.

(*iii*) Each query was translated to its most probable English translation using the probabilistic lexicon constructed with the training part of Agenda 21.

(*iv*) The English query was used to retrieve documents from the English database.

(*v*) The retrieved Dutch documents, together with the Dutch translations of the retrieved English documents, were presented to the user. The user had to decide of each retrieved document, if it was relevant or not.

The volunteers were only confronted with Dutch queries and with Dutch documents that they retrieved. Therefore, their ability to read, write, or translate from, English (or any human's ability to translate from English) did not effect the experiment.

A total of 41 Dutch queries was formulated by the eight volunteers. The experiment shows that even a simple probabilistic lexicon is useful in bilingual document retrieval. With a precision 67% and relative recall of 82%, the bilingual retrieval system surprisingly seems to perform even better than the monolingual Dutch retrieval system, that retrieved documents with a precision of 78%, but relative recall of only 51%.

In order to find an explanation of the unexpected high relative recall of the bilingual retrieval system, the English queries that were generated with the bilingual lexicon were inspected. To each of the English queries a category was assigned according to the following criteria. If the query was a correct translation of the Dutch query, then it was assigned to the *correct* category. If the query was translated incorrect, but was able to retrieve correct fragments the query was assigned to the *usable* category. If the query was translated incorrect because only a part of the original was translated, then it was assigned to the *partially usable* category. If the query could not be translated at all, because the words in the query were not present in the lexicon, then it was assigned to the *not translated* category. Finally, we assigned the remaining queries to the *incorrect* category.

correct	
Dutch query:	*afspraken over samenwerking tussen landen*
translated as:	*arrangements on cooperation between countries*
usable	
Dutch query:	*gezondheid van de mens*
translated as:	*health the human*
partially usable	
Dutch query:	*verbeteren van milieubescherming*
translated as:	*improve protection*
not translated	
Dutch query:	*het kappen van regenwouden in de Filipijnen*
translated as:	? ? in the ?
incorrect	
Dutch query:	*het aandeel van windenergie tot het totaal van energiebronnen*
translated as:	*giving* ? *irrigated energy*

Table 10: translation examples

Of the 41 queries 19 fell into the correct category, 3 fell into the usable category, 10 fell into the partially usable category, 6 fell into the not translated category and 6 fell into the incorrect category. In a Boolean retrieval system, that uses word-based indexing, a translated sentence that is in any of the first two categories (correct or usable) represents a reasonable translation. By this criterion the system performed successfully 54% of the time. Only 6 out of 41 is 15% of the queries were translated incorrect.

4 Conclusion

The research presented in this paper proofs that, as long as a parallel corpus is available on the right domain, it is relatively simple to build a bilingual retrieval system.

The reasons for the unexpected high recall of the bilingual retrieval system, compared to the recall of the monolingual retrieval system can be found in table 10. Queries that fell into the *partially usable* category often contained Dutch compounds (see table 9) that ought to be translated into two separate English words. Our translation model is only able to find one of these words, possibly increasing the recall, but decreasing the precision of the system. The reason that this often leads to an improvement of the recall is the limitation of our domain and the limitation of our corpus. For example, the query *armoedebestrijding* (i.e. *combating poverty*) is, because it is a Dutch compound, translated to *poverty*. However, if we are talking about poverty in the domain of Agenda 21, we usually talk about the combating of poverty. If our database contained fragments of other domains, the recall would not be increased as much as it did now.

Still, the Dutch compound nouns do not explain why the recall of bilingual retrieval was so much higher than the recall of monolingual retrieval. It seems that there is a more structural reason. The lexicon we derived consisted of 3854 English words and 5462 Dutch words. The difference can be explained by Dutch compounds, but also by the use of more synonyms in Dutch and by the richer Dutch morphology (see table 8a). These linguistic phenomena of Dutch, make monolingual Dutch document retrieval a more complicated challenge than monolingual English document retrieval. More research is therefore needed, not only on better translation models, but also on the performance of information retrieval techniques on languages other than English.

References

[1] P.F. Brown, J.C. Lai and R.L. Mercer: "Aligning sentences in parallel corpora", *Proceedings of the 29th Annual Meeting for the Association for Computational Linguistics*, Berkeley CA, 1991, pp. 169-176

[2] W.A. Gale and K.W. Church: "A Program for Aligning Sentences in Bilingual Corpora", *Computational Linguistics*, vol.19(1), 1993, pp.75-102

[3] M. Kay and M. Röscheisen: "Text-translation alignment", *Computational Linguistics*, vol.19(1), 1993, pp. 121-142

[4] A.P. Dempster, N.M. Laird and D.B. Rubin: "Maximum Likelihood from Incomplete Data via the EM-algorithm" plus "discussions on the paper", *Journal of the Royal Statistical Society*, vol.39(B), 1977, pp. 1-38

[5] P.F. Brown, J.C. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin: "A Statistical Approach to Machine Translation", *Computational Linguistics*, vol.16(2), 1990, pp. 79-85

[6] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer: "The Mathematics of Statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, vol.19(2), 1993, pp. 263-313

[7] R. Kruithof, *De Ingenieur*, vol.52(E), 1937, pp.15-25

[8] D. Hiemstra: "Using Statistical Methods to create a Bilingual Dictionary", *Master's thesis, University of Twente*, 1996