Procesamiento del Lenguaje Natural, nº39 (2007), pp. 249-255 recibido 18-05-2007; aceptado 22-06-2007

Flexible statistical construction of bilingual dictionaries

Ismael Pascual Nieto
Universidad Autónoma de Madrid
Escuela Politécnica Superior
ismael.pascual@uam.es

Mick O'Donnell
Universidad Autónoma de Madrid
Escuela Politécnica Superior
michael.odonnell@uam.es

Resumen: La mayoría de los sistemas previos para construir un diccionario bilingüe a partir de un corpus paralelo dependen de un algoritmo iterativo, usando probabilidades de traducción de palabras para alinear palabras en el corpus y sus alineamientos para estimar probabilidades de traducción, repitiendo hasta la convergencia. Si bien este enfoque produce resultados razonables, es computacionalmente lento, limitando el tamaño del corpus que se puede analizar y el del diccionario producido. Nosotros proponemos una aproximación no iterativa para producir un diccionario bilingüe unidireccional que, si bien menos precisa que las aproximaciones iterativas, es mucho más rápida, permitiendo procesar córpora mayores en un tiempo razonable. Asimismo, permite una estimación en tiempo real de la probabilidad de traducción de un par de términos, lo que significa que permite obtener un diccionario de traducción con los *n* términos más frecuentes, y calcular las probabilidades de traducción de términos infrecuentes cuando se encuentren en documentos reales.
Palabras clave: diccionarios bilingües, modelos palabra-a-palabra, traducción automática estadística

Abstract: Most previous systems for constructing a bilingual dictionary from a parallel corpus have depended on an iterative algorithm, using word translation probabilities to align words in the corpus, and using word alignments to estimate word translation probabilities, and repeating until convergence. While this approach produces reasonable results, it is computationally slow, limiting the size of the corpus that can be analysed and the size of the dictionary produced. We propose a non-iterative approach for producing a uni-directional bilingual dictionary which, while less accurate than iterative approaches, is far quicker, allowing larger corpora to be processed in reasonable time. The approach also allows *on-the-fly* estimation of translation likelihoods between a pair of terms, meaning that a translation dictionary can be generated with the *n* most frequent terms in an initial pass, and the translation likelihood of infrequent terms can be calculated as encountered in real documents.
Keywords: bilingual dictionaries, word-to-word models, statistical machine translation

1 *Introduction*

Over the last 17 years, statistical models have been used to construct bilingual dictionaries from parallel corpora, with the goal of using the dictionaries for tasks such as Machine Translation or Cross-Lingual Information Retrieval.

Most of these works have involved an iterative method to construct the dictionary, which start with an initial estimate of word translation probability, use these probabilities to align the words of the corpus, and then use the word alignments to re-estimate word translation probability. This approach cycles until convergence. Followers of this approach include Brown et al. (1990) Kay and Röscheisen, (1993); Hiemstra, (1996); Melamed, (1997); Renders et al., (2003) and Tufis, (2004).

However, the iterative approach is expensive in computing time, requiring extensive calculations on each iteration. Due to memory limitations, these approaches usually restrict consideration to the *n* most frequent terms in each language.

In this paper, we propose a non-iterative approach to building a uni-directional

ISSN: 1135-5948

translation dictionary. While our approach initially produces dictionaries with lower precision, this should be seen in relation to the reduced time needed to build the dictionary. Additionally, our approach supports on-the-fly calculation of the translation suitability between a pair of words. When aligning words in two sentences and less frequent words are encountered, an estimate of the translation likelihood can be derived on the spot, avoiding the need to pre-calculate all possible translation likelihoods between the 76,000 unique terms in our English corpus and the 130,000 unique terms in our Spanish corpus.

The paper is organized as follows: Section 2 discusses the most representative iterative approaches. Section 3 and 4 describes our corpus, and how it is compiled into a word lookup table. Section 5 describes the derivation of our translation dictionaries. Section 6 evaluates the precision and recall of each of our models. Section 7 presents our conclusions.

2 Iterative Approaches

The first published work outlining the construction of bilingual dictionaries using statistical methods was in Brown et al. (1990)[1] at IBM. They used 40,000 aligned sentences from the Canadian HANSARDs corpus (parliament transcripts in English and French).

In their approach, the translation probability between any pair of words is initially set as equi-probable, as are the probabilities of each relative sentence position of a word and its translation. These probabilities are then used to estimate the probability of each possible alignment of the words in each sentence-pair. These probability-weighted alignments are then used to re-estimate the word-translation probabilities as well as the relative position probabilities. This approach cycles until convergence occurs. They used the Expectation Maximization (EM) algorithm.

Subsequent investigators found the IBM approach too computationally complex (requiring iterative re-estimation of 81 million parameters), and the approach did not scale up to larger parallel corpora. Various approaches were tried to improve the performance.

Hiemstra (1996) attempted to reduce complexity using a modified version of the EM algorithm. While the goal of the IBM work was a unidirectional dictionary, Hiemstra aimed to compile a bi-directional dictionary. Hiemstra claimed that the use of bidirectional dictionaries not only reduces the space needed for dictionary storage, but leads to better estimates of translation probabilities[2]. His results improve on those of IBM.

Melamed (1997) proposed an alternative approach, which, while still iterative, required the estimation of fewer parameters. Like the IBM team, he used the HANSARDs corpus, although using 300,000 aligned sentences. He reports 90% precision in real domains.

A key concept in these models is the term *co-occurrence*: two tokens *u* and *v* co-occur if *u* appears in one part of an aligned sentence pair and *v* appears in the other part.

In Melamed's model, co-occurrence is estimated through likelihood ratios, $L(u,v)$, each of which represents the likelihood that *u* and *v* are mutual translations. The process estimating these ratios is as follows:

1) Provide an initial estimate of $L(u,v)$ using their co-occurrence frequencies.
2) Use the estimate of $L(u,v)$ to align the words in the matched sentences of the parallel corpus.
3) Build a new estimate of $L(u,v)$ using the word alignments from step (2).
4) Repeat steps (2) and (3) until convergence occurs (no or little change on each cycle).

Melamed aligns the terms in matched sentences using a *competitive linking algorithm*, which basically orders the $L(u,v)$ values in descending order, and taking these values in turn, links the u and v terms in aligned sentences. Linked terms are then disqualified from linking with other relations.

This process also keeps count of the number of links made between each *u, v* pair, and these counts are used to re-estimate $L(u,v)$.

3 Our corpus

We used the EUROPARL corpus (Kohen, 2005), consisting of transcripts of sessions of the European Parliament between 1996 and 2003. Each transcript is provided in 11 languages. These transcripts are generally constructed by translators, as each speaker speaks in their native language. We used only the English and Spanish sections of the corpus.

[1] The first work of IBM on this was 1988, but it was quite preliminar.

[2] This reference is not in the reference list.

The corpus does not come in sentence aligned form, although each transcript is organised into speaker turns. We wrote software to align the sentences within each speaker turn, based on sequence in the turn, and also on approximate correspondence in number of words, similar to the approach of Gale and Church (1993). Sentences which could not be aligned were discarded. This gave us 730,191 correctly aligned sentences, roughly 20 million words in each language.

4 Compiling a Word Occurrence Index

One of our goals was to allow rapid calculation of translation likelihood between any two terms on the fly. This would not be possible if the entire 40 million word corpus had to be processed each time.

To alleviate this problem, we re-compiled the corpus into an index such as used by web search engines: a file is created for each unique token, detailing each occurrence of the token: the file-id (2 bytes) and sentence-id (2 bytes) of the hit, the position of the token within the sentence (1 byte), and the number of terms in the sentence.

Once the index is compiled, it is possible to derive various statistics rapidly. The frequency of a token can be calculated quickly by dividing the file size by 6 (the record size). The relative co-occurrence of an English and Spanish term can be calculated solely by comparing the index files for those two terms. This allows us to calculate the relative co-occurrence between two terms on the fly, if we need to, rather than having to process the entire corpus to find such a result.

Kay and Röscheisen (1993) also build a word lookup index, but only store the sentence id.

5 Compiling the Bilingual Dictionary

Melamed uses word co-occurrence scores only as an initial estimate of translation suitability. For our purposes, we have found that this initial estimate, if handled properly, provides adequate accuracy for many tasks, without the required expense of the iterative recalculation of translation probabilities through a word alignment process. Our likelihood formula is similar to that of Melamed's although modified to allow our method to work on the fly.

Melamed's initial estimate of translation likelihood of a source term *u* as a target term *v* is the ratio of the joint probability of *u* and *v* and the product of the marginal probabilities of *u* and *v*, as can be seen in equation 1.

$$L(u,v) = \frac{P(u,v)}{P(u) \cdot P(v)} \qquad (1)$$

Basically, if *u* and *v* are not related, this ratio should approach 1.0. The stronger the co-occurrence between *u* and *v*, the higher the *L* value. Substituting in estimates for the probabilities, the formula can be re-expressed as equation 2:

$$L(u,v) = \frac{n(u,v)}{n(u) \cdot n(v)} \cdot N \qquad (2)$$

where, $n(u,v)$ is the co-occurrence frequency of *u,v*, N is the total number of co-occurrences and $n(u)$ is the marginal frequency of *u*, calculated as shown in equation by:

$$n(u) = \sum_{v} n(u,v) \qquad (3)$$

5.1 Our Basic model

The inclusion of $n(u)$ and $n(v)$ in Melamed's formula basically require all values for all *u* and *v* to be calculated at the same time, which means one must decide beforehand which terms will be included in the process. This excludes the calculation of likelihood values for other terms encountered while processing text, which is one of our goals.

We thus use a modified formula which can calculate the translation likelihood between a given *u* and a given *v* independently of other terms. Rather than asking what percent of all *co-occurrences* involve *u* and *v*, we ask what percent of *sentence pairs* contain *u* and *v*. In our approach, $P(u,v)$ represents the probability that *u* occurs in a source sentence while *v* appears in a target sentence. $P(u)$ is the probability that *u* will appear in a source sentence, and $P(v)$ is the probability that *v* will appear in the target sentence.

The important point here is that we can now estimate $L(u,v)$ solely by looking at occurrences of a given *u* and *v*, without needing to consider the whole range of possible *u/v* co-occurrences.

A second change from Melamed's approach is that we desire a unidirectional dictionary. For this reason, we instead use formula 4:

$$L(v \mid u) = \frac{P(v \mid u)}{P(v)} \qquad (4)$$

where $P(v|u)$ is the probability of encountering v in a target sentence if u is in the source sentence, and $P(v)$ is the probability of encountering v in a target sentence. As with Melamed's formula, if u and v are unrelated, the L value will approach 1.0, and higher values indicate a relation between them. A value of 2.0 indicates that v is twice as likely to occur if u is in the corresponding sentence.

Given this simplification, we can calculate $L(u,v)$ as follows:

$$P(v \mid u) = \frac{n_S(u,v)}{n_S(v)} \qquad (5)$$

$$P(v) = \frac{n_S(v)}{S} \qquad (6)$$

$$L(v \mid u) = \frac{n_S(u,v) \big/ n_S(u)}{n_S(v) \big/ S} \qquad (7)$$

$$L(v \mid u) = \frac{n_S(u,v)}{n_S(v) \cdot n_S(u)} \cdot S \qquad (8)$$

where $n_s(u,v)$ is the count of sentence-pairs containing both u and v, $n_s(u)$ is the count of sentence-pairs in which the source sentence contains u, $n_s(v)$ is the count of sentence-pairs in which the target sentence contains v, and S is the total sentence count.

We make one further simplification to allow faster calculation. Because only a small percent of sentences will contain the same word more than once, in the general case, the frequency of a word, $n_w(u)$, will be quite close to $n_S(u)$. Similarly, $n_w(v)$ will approximate $n_S(v)$. We thus use $n_w(u)$ and $n_w(v)$ in place of $n_S(u)$ and $n_S(v)$.

The advantage of this approach is that the frequency of each term is readily available: the size of the index file for the term divided by the record length.

We also choose to use $n(u,v)$ to estimate $n_S(u,v)$ and thus count the co-occurrences of u and v in sentence pairs. This statistic can be derived by scanning through the hit files for u and v, counting cases where the terms appear in the same sentence pair.

For efficiency reasons, we initially compute the values of $L(u,v)$ for the 5000 most frequent tokens in English and Spanish. Any value less than 2.0 is dropped.

We heuristically translate this co-occurrence metric to a translation probability by assuming that the probability of u being translated as v is proportional to the size of the L value. Thus, for each English term u, we collect all the Spanish terms v which were not eliminated, and sum their L values, and divide each by the sum, using this as the translation probability of the term.

Table 1 shows the highest 9 alternatives for *absolutely* (another 16 were included in the list). Several of the Spanish terms (shown in italic) are present due to intra-language collocation between *absolutely* and *essential*, *indispensable* or *crucial* (the indirect association problem mentioned by Melamed). Removing these entries will be discussed below.

English	Spanish	$L(v\|u)$	Prob
absolutely	absolutamente	125.50	0.33
absolutely	absoluta	26.67	0.07
absolutely	*imprescindible*	19.75	0.05
absolutely	absoluto	19.18	0.05
absolutely	*indispensable*	16.08	0.04
absolutely	*crucial*	10.84	0.03
absolutely	totalmente	10.77	0.03
absolutely	*esencial*	9.41	0.03
absolutely	*increíble*	9.29	0.03

Table 1: Translation dictionary alternatives

5.2 Adjusted model

A problem arises with the above formula when a term v *nearly always* occurs with term u. If this is the case, $P(v|u)$ will approach $P(v)$, and the L value will approach 1.0.

For this reason, we introduced the slightly modified formula 9 for likelihood, which instead contrasts those cases where v occurs with u against those cases where v occurs without u:

$$L(v \mid u) = \frac{P(v \mid u)}{P(v \mid \neg u)} \qquad (9)$$

This basically magnifies the likelihood values, as previously the denominator was diluted by cases where u and v co-occur. However, the same interpretation is still valid:

if *u* and *v* are not related, the ratio will approach 1.0, while the stronger the correlation, the higher the likelihood value.

5.3 Using relative distance

By looking at translations between European languages, it is easy to see that a source term tends to appear in a similar relative position in its sentence than its translation in the target sentence.

The probabilistic model of Brown et al. (1990) takes into account that a term in position *i* in a source sentence will translate as a term in position *j* in the target sentence with a given probability, conditioned by the length of the two sentences (*l* and *m*). These calculations however depend on an iterative method, which we are avoiding. It also requires large amounts of data to obtain realistic estimates for possible values of *i*, *j*, *l* and *m*.

We thus proposed a simple heuristic to account for the relative position between two terms. We penalise word co-occurrences in relation to the relative distance between the words in their respective sentences. Firstly, given that the source and target sentences may vary in length, we normalise the position of the term in the sentence by dividing its position by the length of the sentence. The relative distance (d_R) between the terms can then be calculated as follows:

$$d_R(i,j) = \left| \frac{i}{l} - \frac{j}{m} \right| \qquad (10)$$

The closer this value is to 0.0 (no relative distance), the more likely that the terms are translations of each other.

When calculating the co-occurrence of a source and target term, rather than just counting 1 each time the terms appear in the same sentence-pair, we discount the increment by subtracting the relative distance between terms, e.g.

$$n_S(v,u) = \sum_{s \in Sp} \sum_{u,v \in s} \left(1 - d_R(pos(u), pos(v))\right) \qquad (11)$$

where *Sp* is the set of aligned sentence pairs and *pos*(*u*) is the absolute position of the term u in the corresponding part of an aligned sentence pair.

Basically, the further the two terms are away from each other, the less it counts as a viable co-occurrence. This heuristic step improves our results, and the calculation is far simpler than that used in the IBM work.

6 *Evaluation*

Using the above methods, we produced four translation dictionaries, using both the basic and adjusted model, both with and without the distance metric.

We then evaluated the quality of these dictionaries against a gold-standard, *G*, a handcrafted dictionary of 50 terms with human-judged translations. The terms were taken from random positions throughout the word frequency list, and covering a range of syntactic classes.

We then used *G* to evaluate each of the four dictionaries. In terms of precision, for each English term in *G*, we collected the correct translations included in our dictionary, and summed their probability estimates. We then averaged the precision over the 50 terms in *G*. Results for the 4 models are shown in Figure 1.

Our basic dictionary contains up to 25 translation candidates for each source term, with the higher ones being more probable. This list is good for some applications (e.g., word alignment), but produces poor precision (69.96% in the best case). Where precision is important, e.g., for machine translation, we can restrict the number of translation candidates. We achieve 91.94% precision if we just consider the top two candidates.

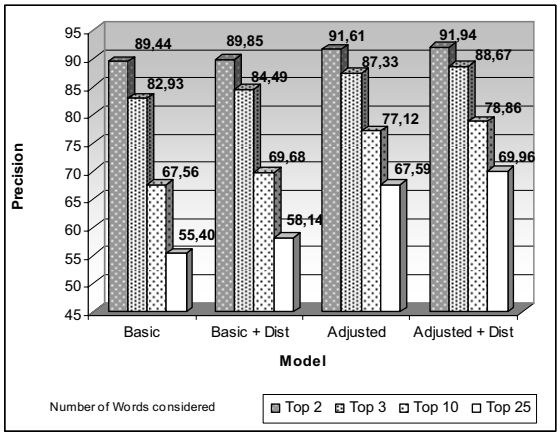

Figure 1: Precision results for the four models

We calculate the recall of a dictionary entry as the percentage of all the correct translations of a term which are in our dictionary. The global recall is then taken as the average over all 50 words. Figure 2 shows our results, again with various levels of cut-off. Our best result was 68.44%, which is quite good considering

many of the translations in the golden standard were not used in the corpus.

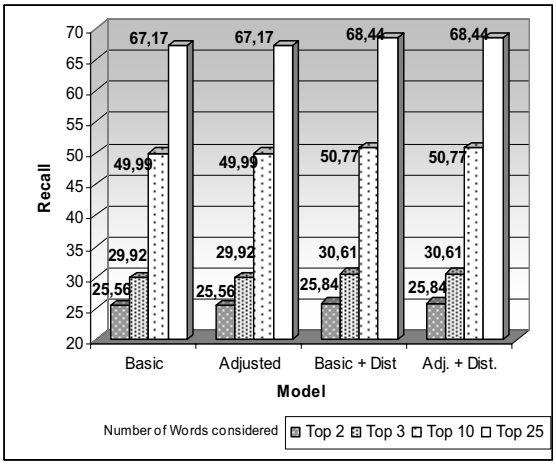

Figure 2: Recall results for the four models

It is clear that including more terms in our dictionary increases recall at the expense of precision. The choice of how many terms to include depends on the application, whether precision or recall is more important.

In terms of assessing which of our 4 models is best, it is clear that the adjusted formula and the inclusion of distance penalties both improve precision, and the distance metric improves recall. Our best model is thus the adjusted+distance one.

6.1 Removing Indirect Associations

One of Melamed's main reasons for taking an iterative approach is to remove false translations due to collocations between source terms. For instance, English *absolutely* is frequently followed by *essential*, and for this reason, *absolutely* has strong co-occurrence with words which translate *essential*.

Melamed only uses co-occurrence values as the basis for aligning words in sentences, and the aligned words are then used to re-estimate word translation probabilities. Since the true translation of a word will generally have a higher co-occurrence value than the false translations, the collocation-induced mappings will be dropped from the data.

One of the prime uses of our translation dictionary is to support word alignment. When used for this purpose, the presence of indirect associations in our dictionary is generally not a problem, because the term with a direct association will be the preferred alignment choice.

However, when using our dictionary for other tasks, such as automatic sentence translation, the indirect associations will be a problem.

For this reason, we have developed a method to remove indirect associations from our dictionary, a means which does not require the expensive step of word-aligning the entire corpus. We firstly derive collocation values between words of the same language. We then pass through our translation dictionary, and whenever a translation of a term is also the translation of a collocate of the term, the co-occurrence value is recalculated, using only those cases where the collocate is not present.

We applied this process as a post-operation on the translation dictionaries produced earlier. Looking only at the adjusted+distance model with 25 translations, removing indirect associations increased precision from 69.96% to 74.85%, a significant increase. Recall also rose from 68.44% to 69.80%. See Figures 3 and 4.

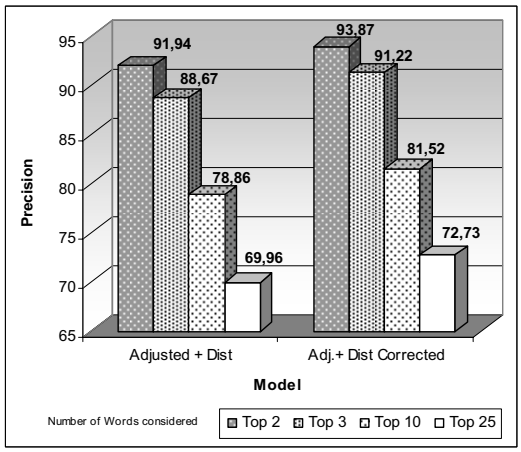

Figure 3: Adjusted+distance model with and without collocation correction: Precision

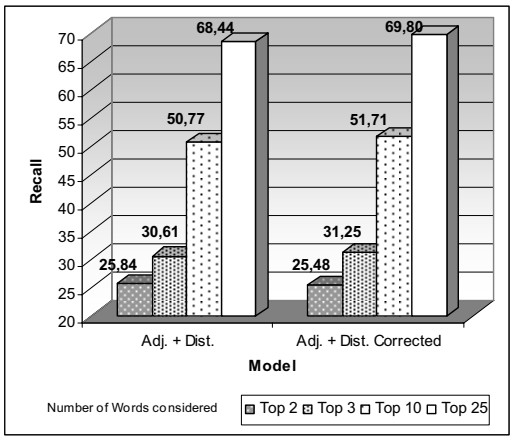

Figure 4: Adjusted+distance model with and without collocation correction: Recall.

7 Conclusions and future work

In this paper, we proposed an approach to building bilingual dictionaries from a parallel corpus which avoids the computational complexity of the iterative approaches. The approach allows calculation of translation likelihood of pair of words without needing to consider other words at the same time, as in Melamed's approach. This makes the approach suitable for on-the-fly estimation of translation likelihood of a pair of words encountered during tasks such as aligning words in parallel sentences.

To avoid the problem of indirect association, we propose a method to eliminate such effects from the likelihood table without needing to word-align the corpus.

While our levels of precision and recall are not as high as the iterative approaches, the speed and flexibility of our approach makes it a viable candidate for cases where computation time is an issue, or where building larger dictionaries in realistic timeframes is required.

In terms of the various models we have experimented with, we found that our adjusted model, using $P(v|u)/P(v|\neg u)$, gave higher precision than the more pure likelihood measure: $P(v|u)/P(v)$. Also, including distance penalties improved both approaches.

References

Brown, P.F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin. 1990. A statistical approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Gale, W.A. and K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Hiemstra, D. 1996. Using statistical methods to create a bilingual dictionary. Master Thesis. University of Twente.

Kay, M., M. Röscheisen. 1993. Text-Translation Alignment. *Computational Linguistics* 19(1): 121-142.

Koehn, P. 2005. Europarl: A parallel corpus for Statistical Machine Translation. In: *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Melamed, I.D. 1997. A word-to-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 490–497

Renders, J.-M., H. Déjean and É. Gaussier. 2003. Assessing automatically extracted bilingual lexicons for CLIR in vertical domains. *Lecture Notes in Computer Science* 2785, C. Peters, M. Braschler, J. Gonzalo and M. Kluck Editors, Springer-Verlag: Berlin, pp. 363–371.

Tufis, D. and A.M. Barbu and R. Ion. 2004. Extracting multilingual lexicons from parallel corpora. *Computers and the Humanities*, 38(2):163–189.