AN ALGORITHM FOR FINDING NOUN PHRASE CORRESPONDENCES IN BILINGUAL CORPORA

Julian Kupiec
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
kupiec@parc.xerox.com

Abstract

The paper describes an algorithm that employs English and French text taggers to associate noun phrases in an aligned bilingual corpus. The taggers provide part-of-speech categories which are used by finite-state recognizers to extract simple noun phrases for both languages. Noun phrases are then mapped to each other using an iterative re-estimation algorithm that bears similarities to the Baum-Welch algorithm which is used for training the taggers. The algorithm provides an alternative to other approaches for finding word correspondences, with the advantage that linguistic structure is incorporated. Improvements to the basic algorithm are described, which enable context to be accounted for when constructing the noun phrase mappings.

INTRODUCTION

Areas of investigation using bilingual corpora have included the following:

- Automatic sentence alignment [Kay and Röscheisen, 1988, Brown *et al.*, 1991a, Gale and Church, 1991b].
- Word-sense disambiguation [Dagan *et al.*, 1991, Brown *et al.*, 1991b, Church and Gale, 1991].
- Extracting word correspondences [Gale and Church, 1991a].
- Finding bilingual collocations [Smadja, 1992].
- Estimating parameters for statistically-based machine translation [Brown *et al.*, 1992].

The work described here makes use of the aligned Canadian Hansards [Gale and Church, 1991b] to obtain noun phrase correspondences between the English and French text.

The term "correspondence" is used here to signify a mapping between words in two aligned sentences. Consider an English sentence E_i and a French sentence F_i which are assumed to be approximate translations of each other. The subscript i denotes the i'th alignment of sentences in both languages. A word sequence in E_i is defined here as the correspondence of another sequence in F_i if the words of one sequence are considered to represent the words in the other.

Single word correspondences have been investigated [Gale and Church, 1991a] using a statistic operating on contingency tables. An algorithm for producing collocational correspondences has also been described [Smadja, 1992]. The algorithm involves several steps. English collocations are first extracted from the English side of the corpus. Instances of the English collocation are found and the mutual information is calculated between the instances and various single word candidates in aligned French sentences. The highest ranking candidates are then extended by another word and the procedure is repeated until a corresponding French collocation having the highest mutual information is found.

An alternative approach is described here, which employs simple iterative re-estimation. It is used to make correspondences between simple noun phrases that have been isolated in corresponding sentences of each language using finite-state recognizers. The algorithm is applicable for finding single or multiple word correspondences and can accommodate additional kinds of phrases.

In contrast to the other methods that have been mentioned, the algorithm can be extended in a straightforward way to enable correct correspondences to be made in circumstances where numerous low frequency phrases are involved. This is important consideration because in large text corpora roughly a third of the word types only occur once.

Several applications for bilingual correspondence information have been suggested. They can be used in bilingual concordances, for automatically constructing bilingual lexicons, and probabilistically quantified correspondences may be useful for statistical translation methods.

COMPONENTS

Figure 1 illustrates how the corpus is analyzed. The words in sentences are first tagged with their corresponding part-of-speech categories. Each tagger contains a hidden Markov model (HMM), which is trained using samples of raw text from the Hansards for each language. The taggers are robust and operate with a low error rate [Kupiec, 1992]. Simple noun phrases (excluding pronouns and digits) are then extracted from the sentences by finite-state recognizers that are specified by regular expressions defined in terms of part-of-speech categories. Simple noun phrases are identified because they are most reliably recognized; it is also assumed that they can be identified unambiguously. The only embedding that is allowed is by prepositional phrases involving "of" in English and "de" in French, as noun phrases involving them can be identified with relatively low error (revisions to this restriction are considered later). Noun phrases are placed in an index to associate a unique identifier with each one.

A noun phrase is defined by its word sequence, excluding any leading determiners. Singular and plural forms of common nouns are thus distinct and assigned different positions in the index. For each sentence corresponding to an alignment, the index positions of all noun phrases in the sentence are recorded in a separate data structure, providing a compact representation of the corpus.

So far it has been assumed (for the sake of simplicity) that there is always a one-to-one mapping between English and French sentences. In practice, if an alignment program produces blocks of several sentences in one or both languages, this can be accommodated by treating the block instead as a single bigger "compound sentence" in which noun phrases have a higher number of possible correspondences.

THE MAPPING ALGORITHM

Some terminology is necessary to describe the algorithm concisely. Let there be L total alignments in the corpus; then E_i is the English sentence for alignment i. Let the function $\phi(E_i)$ be the number of noun phrases identified in the sentence. If there are k of them, $k = \phi(E_i)$, and they can be referenced by $j = 1...k$. Considering the j'th noun phrase in sentence E_i, the function $\mu(E_i, j)$ produces an identifier for the phrase, which is the position of the phrase in the English index. If this phrase is at position s, then $\mu(E_i, j) = s$.

In turn, the French sentence F_i will contain $\phi(F_i)$ noun phrases and given the p'th one, its position in the French index will be given by $\mu(F_i, p)$. It will also be assumed that there are a total of V_E and V_F phrases in the English and French indexes respectively. Finally, the indicator function $I()$ has the value unity if its argument is true, and zero otherwise.

Assuming these definitions, the algorithm is

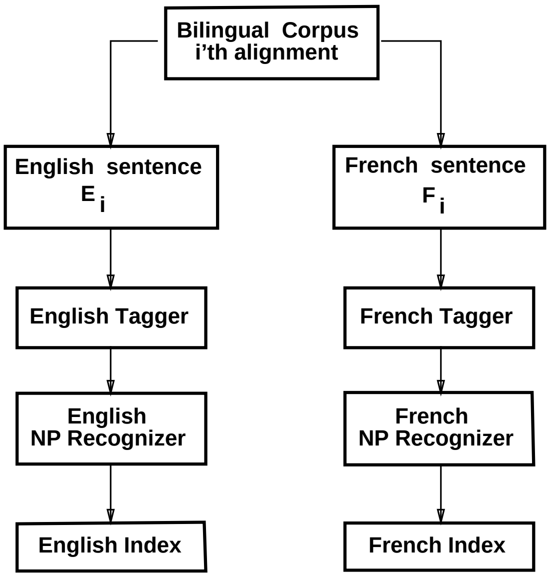

Figure 1: Component Layout

stated in Figure 2. The equations assume a directionality: finding French "target" correspondences for English "source" phrases. The algorithm is reversible, by swapping E with F.

The model for correspondence is that a source noun phrase in E_i is responsible for producing the various different target noun phrases in F_i with correspondingly different probabilities. Two quantities are calculated; $C_r(s,t)$ and $P_r(s,t)$. Computation proceeds by evaluating Equation (1), Equation (2) and then iteratively applying Equations (3) and (2); r increasing with each successive iteration. The argument s refers to the English noun phrase $np_E(s)$ having position s in the English index, and the argument t refers to the French noun phrase $np_F(t)$ at position t in the French index. Equation (1) assumes that each English noun phrase in E_i is initially equally likely to correspond to each French noun phrase in F_i. All correspondences are thus equally weighted, reflecting a state of ignorance. Weights are summed over the corpus, so noun phrases that co-occur in several sentences will have larger sums. The weights $C_0(s,t)$ can be interpreted as the mean number of times that $np_F(t)$ corresponds to $np_E(s)$ given the corpus and the initial assumption of equiprobable correspondences.

These weights can be used to form a new estimate of the probability that $np_F(t)$ corresponds to $np_E(s)$, by considering the mean number of times $np_F(t)$ corresponds to $np_E(s)$ as a fraction of the total mean number of correspondences for $np_E(s)$, as in Equation (2). The procedure is then iterated using Equations (3), and (2) to obtain successively refined, convergent estimates of the prob-

$$\begin{aligned}
C_0(s,t) &= \sum_{i=1}^{L}\sum_{j=1}^{\phi(E_i)}\sum_{k=1}^{\phi(F_i)} I(\mu(E_i,j)=s)I(\mu(F_i,k)=t)\frac{1}{\phi(F_i)} && (1)\\
P_r(s,t) &= \frac{C_{r-1}(s,t)}{\sum_{q=1}^{V_F} C_{r-1}(s,q)} && (2)\\
C_r(s,t) &= \sum_{i=1}^{L}\sum_{j=1}^{\phi(E_i)}\sum_{k=1}^{\phi(F_i)} I(\mu(E_i,j)=s)I(\mu(F_i,k)=t)P_{r-1}(s,t) && (3)\\
r &> 0\\
V_E \geq s &\geq 1\\
V_F \geq t &\geq 1
\end{aligned}$$

Figure 2: The Algorithm

ability that $np_F(t)$ corresponds to $np_E(s)$. The probability of correspondences can be used as a method of ranking them (occurrence counts can be taken into account as an indication of the reliability of a correspondence). Although Figure 2 defines the coefficients simply, the algorithm is not implemented literally from it. The algorithm employs a compact representation of the correspondences for efficient operation. An arbitrarily large corpus can be accommodated by segmenting it appropriately.

The algorithm described here is an instance of a general approach to statistical estimation, represented by the EM algorithm [Dempster *et al.*, 1977]. In contrast to reservations that have been expressed [Gale and Church, 1991a] about using the EM algorithm to provide word correspondences, there have been no indications that prohibitive amounts of memory might be required, or that the approach lacks robustness. Unlike the other methods that have been mentioned, the approach has the capability to accommodate more context to improve performance.

RESULTS

A sample of the aligned corpus comprising 2,600 alignments was used for testing the algorithm (not all of the alignments contained sentences). 4,900 distinct English noun phrases and 5,100 distinct French noun phrases were extracted from the sample.

When forming correspondences involving long sentences with many clauses, it was observed that the position at which a noun phrase occurred in E_i was very roughly proportional to the corresponding noun phrase in F_i. In such cases it was not necessary to form correspondences with all noun phrases in F_i for each noun phrase in E_i. Instead, the location of a phrase in E_i was mapped linearly to a position in F_i and correspondences were formed for noun phrases occurring in a window around that position. This resulted in a total of 34,000 correspondences. The mappings are stable within a few (2-4) iterations.

In discussing results, a selection of examples will be presented that demonstrates the strengths and weaknesses of the algorithm. To give an indication of noun phrase frequency counts in the sample, the highest ranking correspondences are shown in Table 1. The figures in columns (1) and (3) indicate the number of instances of the noun phrase to their right.

185	Mr. Speaker	187	M. Le Président
128	Government	141	gouvernement
60	Prime Minister	65	Premier Ministre
63	Hon. Member	66	député
67	House	68	Chambre

Table 1: Common correspondences

To give an informal impression of overall performance, the hundred highest ranking correspondences were inspected and of these, ninety were completely correct. Less frequently occurring noun phrases are also of interest for purposes of evaluation; some of these are shown in Table 2.

32	Atlantic Canada Opportunities Agency	23	Agence de promotion économique du Canada atlantique
5	DREE	4	MEER
1	late spring	1	fin du printemps
1	whole issue of free trade	1	question du libre-échange

Table 2: Other correspondences

The table also illustrates an unembedded English noun phrase having multiple prepositional phrases in its French correspondent. Organizational acronyms (which may be not be available in general-purpose dictionaries) are also extracted, as the taggers are robust. Even when a noun phrase only occurs once, a correct correspondence can be found if there are only single noun phrases in each sentence of the alignment. This is demonstrated in the last row of Table 2, which is the result of the following alignment:

E_i: "The whole issue of free trade has been mentioned."

F_i: "On a mentionné la question du libre-échange."

Table 3 shows some incorrect correspondences produced by the algorithm (in the table, "usine" means "factory").

1	off-the-job training	6	usine
1	mix of on-the-job	6	usine

Table 3

The sentences that are responsible for these correspondences illustrate some of the problems associated with the correspondence model:

E_i: "They use what is known as the dual system in which there is a mix of on-the-job and off-the-job training."

F_i: "Ils ont recours à une formation mixte, partie en usine et partie hors usine."

The first problem is that the conjunctive modifiers in the English sentence cannot be accommodated by the noun phrase recognizer. The tagger also assigned "on-the-job" as a noun when adjectival use would be preferred. If verb correspondences were included, there is a mismatch between the three that exist in the English sentence and the single one in the French. If the English were to reflect the French for the correspondence model to be appropriate, the noun phrases would perhaps be "part in the factory" and "part out of the factory". Considered as a translation, this is lame. The majority of errors that occur are not the result of incorrect tagging or noun phrase recognition, but are the result of the approximate nature of the correspondence model. The correspondences in Table 4 are likewise flawed (in the table, "souris" means "mouse" and "tigre de papier" means "paper tiger"):

1	toothless tiger	1	souris
1	toothless tiger	1	tigre de papier
1	roaring rabbit	1	souris
1	roaring rabbit	1	tigre de papier

Table 4

These correspondences are the result of the following sentences:

E_i: "It is a roaring rabbit, a toothless tiger."

F_i: "C' est un tigre de papier, un souris qui rugit."

In the case of the alliterative English phrase "roaring rabbit", the (presumably) rhetorical aspect is preserved as a rhyme in "souris qui rugit"; the result being that "rabbit" corresponds to "souris" (mouse). Here again, even if the best correspondence were made the result would be wrong because of the relatively sophisticated considerations involved in the translation.

EXTENSIONS

As regards future possibilities, the algorithm lends itself to a range of improvements and applications, which are outlined next.

Finding Word Correspondences: The algorithm finds corresponding noun phrases but provides no information about word-level correspondences within them. One possibility is simply to eliminate the tagger and noun phrase recognizer (treating all words as individual phrases of length unity and having a larger number of correspondences). Alternatively, the following strategy can be adopted, which involves fewer total correspondences. First, the algorithm is used to build noun phrase correspondences, then the phrase pairs that are produced are themselves treated as a bilingual noun phrase corpus. The algorithm is then employed again on this corpus, treating all words as individual phrases. This results in a set of single word correspondences for the internal words in noun phrases.

Reducing Ambiguity: The basic algorithm assumes that noun phrases can be uniquely identified in both languages, which is only true for simple noun phrases. The problem of prepositional phrase attachment is exemplified by the following correspondences:

16	Secretary of State	20	secrétaire d' Etat
16	Secretary of State	19	Affaires extérieures
16	External Affairs	19	Affaires extérieures
16	External Affairs	20	secrétaire d' Etat

Table 5

The correct English and French noun phrases are "Secretary of State for External Affairs" and "secrétaire d' Etat aux Affaires extérieures". If prepositional phrases involving "for" and "à" were also permitted, these phrases would be correctly identified; however many other adverbial prepositional phrases would also be incorrectly attached to noun phrases.

If all embedded prepositional phrases were permitted by the noun phrase recognizer, the algorithm could be used to reduce the degree of ambiguity between alternatives. Consider a sequence $np_e pp_e$ of an unembedded English noun phrase np_e followed by a prepositional phrase pp_e, and likewise a corresponding French sequence $np_f pp_f$. Possible interpretations of this are:

1. The prepositional phrase attaches to the noun phrase in both languages.
2. The prepositional phrase attaches to the noun phrase in one language and does not in the other.
3. The prepositional phrase does not attach to the noun phrase in either language.

If the prepositional phrases attach to the noun phrases in both languages, they are likely to be repeated in most instances of the noun phrase; it is less likely that the same prepositional phrase will be used adverbially with each instance of the noun phrase. This provides a heuristic method for reducing ambiguity in noun phrases that occur several times. The only modifications required to the algorithm are that the additional possible noun phrases and correspondences between them must be included. Given thresholds on the number of occurrences and the probability of the correspondence, the most likely correspondence can be predicted.

Including Context: In the algorithm, correspondences between source and target noun phrases are considered irrespectively of other correspondences in an alignment. This does not make the best use of the information available, and can be improved upon. For example, consider the following alignment:

E_i: "The Bill was introduced just before Christmas."

F_i: "Le projet de loi a été presenté juste avant le congé des Fêtes."

Here it is assumed that there are many instances of the correspondence "Bill" and "projet de loi", but only one instance of "Christmas" and "congé des Fêtes". This suggests that "Bill" corresponds to "projet de loi" with a high probability and that "Christmas" likewise corresponds strongly to "congé des Fêtes". However, the model will assert that "Christmas" corresponds to "projet de loi" and to "congé des Fêtes" with equal probability, no matter how likely the correspondence between "Bill" and "projet de loi".

The model can be refined to reflect this situation by considering the joint probability that a target $np_F(t)$ corresponds to a source $np_E(s)$ *and all the other possible correspondences in the alignment are produced.* This situation is very similar to that involved in training HMM text taggers, where joint probabilities are computed that a particular word corresponds to a particular part-of-speech, and the rest of the words in the sentence are also generated (e.g. [Cutting *et al.*, 1992]).

CONCLUSION

The algorithm described in this paper provides a practical means for obtaining correspondences between noun phrases in a bilingual corpus. Linguistic structure is used in the form of noun phrase recognizers to select phrases for a stochastic model which serves as a means of minimizing errors due to the approximations inherent in the correspondence model. The algorithm is robust, and extensible in several ways.

References

[Brown *et al.*, 1991a] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, pages 169–176, Berkeley, CA., June 1991.

[Brown *et al.*, 1991b] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, pages 264–270, Berkeley, CA., June 1991.

[Brown *et al.*, 1992] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. D. Lafferty, and R. L. Mercer. Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 83–100, Montreal, Canada., June 1992.

[Church and Gale, 1991] K. W. Church and W. A. Gale. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research*, pages 40–62, September 1991.

[Cutting *et al.*, 1992] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, April 1992. ACL.

[Dagan *et al.*, 1991] I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association of Computational*

Linguistics, pages 130–137, Berkeley, CA., June 1991.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[Gale and Church, 1991a] W. A. Gale and K. W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 152–157, Pacific Grove, CA., February 1991. Morgan Kaufmann.

[Gale and Church, 1991b] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, pages 177–184, Berkeley, CA., June 1991.

[Kay and Röscheisen, 1988] M. Kay and M. Röscheisen. Text-translation alignment. Technical Report P90-00143, Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304, June 1988.

[Kupiec, 1992] J. M. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech and Language*, 6:225–242, 1992.

[Smadja, 1992] F. Smadja. How to compile a bilingual collocational lexicon automatically. In C. Weir, editor, *Proceedings of the AAAI-92 Workshop on Statistically-Based NLP Techn2----hrase*