[TMI 99: *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, August 1999, Chester, UK]

A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns

Kevin McTait, Arturo Trujillo
Centre for Computational Linguistics
UMIST
Manchester, PO BOX 88, M60 1QD, UK
{kevimn,iat}@ccl.umist.ac.uk

Abstract

In this paper, we present an algorithm for the automatic extraction of *translation patterns* between two (Indo-)European languages. These consist of possibly discontiguous text fragments, with the bilingual relationship between the text fragments and the discontinuities between them made explicit. The patterns are extracted from a bilingual parallel corpus aligned at the sentence level, without the need for linguistic analysis, and are used to build a translation memory database which is intended for use in a machine aided human translation (MAHT) setting, such as a translator's workbench (TWB). The patterns extracted could also form the basis for example-based machine translation (EBMT) without the need for complex linguistic or statistical processing. Given a TM database made up of our concept of translation patterns and a SL input string, relevant translation patterns combine to form TL translations as suggestions to the translator. We evaluate the accuracy of the translation patterns extracted along with the quality of translations produced.

1 Introduction

Traditional Translation Memory (TM) systems make use of a database of sentences that are translations of each other. Given a SL input string to translate, the system has to find the best possible match within a very specific set of data. By contrast, the related task of EBMT requires much linguistic or statistical pre-processing including tagging and parsing in order to process and extract suitable examples. These two approaches lie at opposite ends of a spectrum in Memory Based Translation, each with their own advantages and disadvantages. What is required is a more flexible TM mechanism which is less knowledge intensive. If we could make the TM database more general, the SL input would be more likely to match sentences and have its target forms generated correctly. This would also provide the basis for shallower and less complex EBMT. In effect, we aim for an approach that lies somewhere between traditional TM and EBMT.

In this paper, we present an algorithm for the automatic extraction of translation patterns from a bilingual corpus aligned at the sentence level. By translation patterns we mean generalisations of sentences that are translations of each other. We generalise these sentential translations by identifying recurring word groups, aligning them and then aligning any *slots* that may occur between these word groups. Such translation patterns are more useful in a TM system than a set of complete sentences that are translations of each other as is traditionally used in commercial TM systems. The output of our algorithm is a set of generalised sentence translations or translation patterns with the relationship between the strings and the variables or slots made explicit, as in (1) below. This example shows how an English sentence containing *gave...up* can be translated by a Spanish sentence containing *abandonó.* In this case, the strings *gave* and *up* are aligned with *abandonó.* The slots stand for a sequence of one or more tokens and are assigned the variables *X* and *Y* to show how they are aligned.

(1) X_s *gave* Y_s *up* ⇔ X_t *abandonó* Y_t

Our algorithm can be divided into 3 clear stages:

1. The Monolingual Phase: The input to this stage is the bilingual corpus aligned at the level of the sentence. Lexical items or tokens that occur in 2 or more sentences are collected. Prom these, all recurrent contiguous and non-contiguous strings (Collocations) are collected, noting where the discontinuities arise. This is done for sentences in both languages.
2. The Bilingual Phase: The alignment of Collocations between languages is based on simple co-occurrence criteria.
3. String and Slot Alignment: Given a set of aligned Collocations, the strings and the discontinuities or slots are aligned to produce translation patterns.

The rationale behind this algorithm is that possibly discontinuous pairs of source and target strings that co-occur in 2 or more sentences are very likely to be translations of each other.

Our algorithm is language independent in nature and operates on the simple principles of string co-occurrence and frequency thresholds. Since we only require strings to co-occur a minimum of twice in the corpus, our algorithm is useful for analysing sparse data. This makes it ideal for 'less-studied' languages where there are limited supplies of bilingual corpora. In addition, the algorithm is incremental. This facilitates the construction of translator resources from scratch for new languages and domains and assists new translators in building translation resources.

2 Background

Our work is closely related to the task of bilingual vocabulary and term alignment. Most of the literature in this field concentrates on aligning single words (Dagan et al. 1993), terms (van der Eijk 1993), single words and terms (Fung & McKeown 1997) and even collocations (Smadja et al. 1996). All of these approaches are based upon monolingually extracting lexical units of varying lengths to be later aligned by means of statistical correlation and/or thresholds, with equivalent units in a target language. Smadja et al. (1996) make use of the Dice coefficient for aligning collocations, van der Eijk (1993) uses an elaboration of the information theoretic Mutual Information score, while Fung & McKeown (1997) compare recency vectors. Somers (1998) attempts to recreate the work of Fung & McKeown (1997) using Levenshtein distance. Brown

(1997) uses a threshold scheme where only pairs of words that satisfy a threshold pass the filter to be subsequently considered translations. Gaussier (1998) models word alignments as a directed connected graph whose edges reflect alignment probabilities: SL terms are extracted and then the alignment probabilities between them and TL items are estimated and then refined by an iterative procedure until a terminal point is reached. Chen et al. (1997) take a different approach by attempting to introduce knowledge sources: a thesaurus and a bilingual dictionary. Their approach relies on the topical clustering of dictionary entries and their translations in order to provide estimates of lexical translation probability needed to compute $Pr(S \mid T)$.

Perhaps the most closely related work, in terms of output, to that of our own is that of Smadja et al. (1996). They attempt to extract and align contiguous and non-contiguous collocations. They use a tool called Xtract (Smadja 1993) in order to extract collocations from an English source text. They make use of an iterative process that finds associated word pairs by identifying frequently occurring words in recurrent positions around a given word. This process is repeated iteratively by finding highly associated words with these word pairs, and subsequently triples and n-tuples, until no more associations can be made. A robust parser facilitates filtering the output to select semantically meaningful collocations. The bilingual task consists of finding, by means of the Dice coefficient, all TL words associated with each SL collocation. In a similar incremental fashion, TL word pairs are formed from this set and are then associated, using the Dice coefficient, with the SL collocation. Highly correlated words are added to the pairs to produce triplets and eventually n-tuples that are correlated with the SL collocation. The corpus is scanned for instances of the TL collocation in order to determine its most consistent word ordering.

Some approaches in the literature are based either on a specific language pair or make use of linguistic tools or heuristics to guide the algorithms and tune them for optimum performance. van der Eijk (1993) and Smadja et al. (1996) use a POS tagger and shallow parser in the monolingual phase of lexical extraction. Brown (1997) uses weights based on similar positions in order to increase the likelihood that a target word is the translation of a source word. Similarly, Simard et al. (1992) use cognates in retrieving anchor points for sentence alignment. We use the term 'language-neutral' here to describe the absence of specific and explicit linguistic information in the algorithm.

Approaches that make use of separate monolingual modules, involving an amount of linguistic pre-processing for extracting terms, are prone to error and this error is subsequently passed on to the remaining (bilingual) modules. Smadja (1993) reports that 10% of terms extracted are not valid terms and van der Eijk (1993) also reports that recall could be increased if the errors incurred by POS tagging and shallow parsing the source text could be eliminated.

Some researchers define the units they want to align. van der Eijk (1993) defines terms as NPs which are identified by recognising certain sequences of POS tags. Ahrenberg et al. (1998) consider single words and multi-word units in independent extraction and alignment phases. Our approach simply identifies recurrent word patterns of any length.

One difference that our approach has over some of the more complex statistical methods of alignment is that we do not use training data to estimate and refine statistical parameters, as is the case in Dagan et al. (1993), for example. Furthermore, we not only align textual fragments, but also any discontinuities that may occur between them.

Methods that have concentrated on automatically extracting translation patterns for EBMT require linguistic analysis, which in turn requires explicit linguistic and language-dependent resources. Consequently, such patterns contain deeper and more complex syntactic descriptions requiring expensive algorithms. Kaji et al. (1992), for example, use a bilingual dictionary and parser to find correspondences at the phrase structure level between two sentences that are translations of each other. These structures are then replaced by variables to produce translation patterns, similar to those in (1), except that the variables contain syntactic and possibly semantic constraints. The translation patterns described in Watanabe (1993) make use of a complex data structure involving a combination of lexical mappings and mappings between dependency structures, as is the case for the pattern-based CFG rules found in Takeda (1996).

Section 3 describes our approach to extracting translation patterns: the monolingual task of collecting and combining lexical items to produce Collocations, how Collocations are aligned and how we align the strings and the slots in Collocations. Section 4 presents the results of applying this algorithm to a set of 3000 aligned sentence pairs and a preliminary indication of the translation quality possible using these patterns in a TM system. Section 5 outlines our proposals to improve performance.

3 Extracting Translation Patterns

The following is an illustrated example of the methodology of the algorithm. Given the corpus in (2):

(2) *1. The Commission gave the plan up ⇔ La Comisión abandonó el plan*
2. Our Government gave all laws up ⇔ Nuestro Govierno abandonó todas las leyes

the items in (3) below are extracted first, since they occur the specified minimum of twice. The integers denote the sentences from which they were retrieved.

gave [1, 2] abandonó [1, 2]
(3) up [1,2]

The items, *gave* and *up,* retrieved from the English side of the corpus are allowed to combine to form the orthographically longer *Collocation* shown in (4), as they both occur in at least 2 sentences.

(4) *(gave)(up)[l, 2]*

A Collocation is a data structure representing strings that co-occur in 2 or more sentences. The integers denote those sentences. Collocations are aligned via the sentence ids in which they occur. For example, (5) shows the Spanish alignment for Collocation (4).

(5) *(...) gave (...) up ⇔ (...) abandonó (...)*

The *Complement* of Collocation (5) is formed using the text found in sentence 1 of the corpus. It is intuitive that if the strings in the Collocation are translations of each other, then the strings occupying the slots, indicated by *(...)* in (5), should also be translations of each other, as illustrated in (6).

(6) *The Commission (...) the plan ⇔ La Comisión (...) el plan*

The strings and the slots in both the Collocation and the Complements are aligned to produce the translation patterns produced by the algorithm. These are given in (7) below. In this case, *gave* and *up* are aligned with *abandonó, the commission* with *la comisión* and *the plan* with *el plan.*

(7) *X_s gave Y_s up ⇔ X_t abandonó Y_t*
The Commission X_s the plan ⇔ La Comisión X_t el plan

We now describe each phase in more detail.

3.1 Monolingual Phase

This stage is applied independently to both the source and target side of the corpus. Single lexical items, or tokens, that occur twice or more are retrieved, together with a record of the sentences in which they were found. Given the corpus in (2), the lexical items retrieved are shown in (3).

Once all the tokens in one language side of the corpus that occur twice or more have been extracted, they are allowed to combine to form longer word combinations (or Collocations, as we term them, since they form some sort of ‘arbitrary and recurrent word pattern’ (Benson 1990)) constrained only by the sentences from which they were retrieved. This set of Collocations combine recursively to form a tree-like data structure. Each Collocation is tested to see if it can combine with the daughters of the root node and if so, recursively with each subsequent daughter, as long as there is an intersection of at least two sentence ids. This enforces string co-occurrence in 2 or more sentences.

Figure la shows how *gave* is added to the root node. Figure 1b shows how *up* is added and allowed to combine with *gave* to form a longer Collocation. The integers denote the sentences from which the lexical items were taken, while the words indicate the lexical items retrieved. Finally, figure lc shows a larger Collocation tree with further combinations if a larger corpus were used.

When new daughters are added to the nodes, orthographically longer Collocations are produced. In doing so, we obtain a tree of Collocations of increasing orthographic length, but of decreasing frequency. Therefore, the leaves become the most informative parts of the tree. From these leaf-collocations only the longest are selected. We choose the longest Collocations because a longer pattern provides more context and hence there is less possibility of ambiguity.

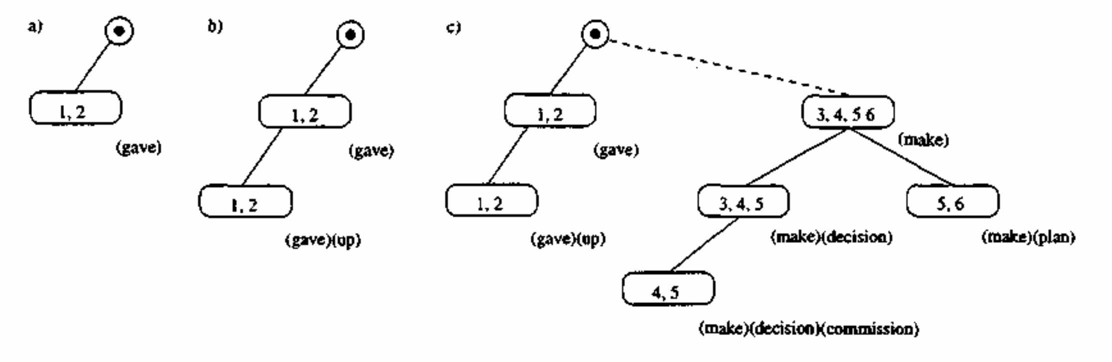

Figure 1: Adding Strings to a Collocation Tree

3.2 Bilingual Phase

Bilingual alignment of Collocations is now a simple task. It uses the simple principle of string co-occurrence in two or more aligned sentences across languages. Collocations from one language side that have exactly the same sentence ids as a Collocation from another language side are considered to be translations of each other, as illustrated in (8). On account of the fact that a Collocation is formed from strings that co-occur in at least 2 sentences, we can choose from any of those sentences and determine the relative word order in the Collocation, as shown in (5). In practice, we choose the word ordering from the sentence that provides us with the least number of slots.

(8) *(gave)(up)[1, 2]*
(abandonó)[1, 2]
(gave)(up) ⇔ *(abandonó)*

As already mentioned, we take the Complement of each possible Collocation. By definition it has at least two. Therefore, for each Collocation, we can return at least two Complements. The Complement is simply made up of the strings that occur between the words that make up a Collocation. The slots in a Collocation thus correspond to the strings in the Complement and vice versa. Given the Collocation in (5) and the sentences from which it was formed, the following Complements are produced (9):

(9) *The Commission (...) the plan (...)* ⇔ *La Comisión (...) el plan*
Our Government (...) all laws (...) ⇔ *Nuestro Govierno (...) todas las leyes*

3.3 Translation Slots

In order to make the translation patterns complete, the strings and the slots must be aligned (figure 2). In doing so, we produce translation patterns similar to those in Langé et al. (1997). The slots in our scheme of translation patterns simply stand for sequences of one or more tokens. Aligning slots is therefore analogous to aligning words, phrases, terms or sentences as in conventional bilingual alignment, and so identifying correspondences between slots involves similar problems (we are in fact aligning sub-sentential components). A knowledge-free approach compounds the problem by not making linguistic information available, such as a bilingual dictionary. While the experiments of Gale & Church (1993) found that aligning sentences in parallel corpora such as the Canadian Hansards revealed relationships such as 1:1, 1:0, 1:2, slot alignment for translation patterns and alignment of sub-sentential components in general has revealed more complex patterns, such as 1:3, 3:4, 4:5, 3:5, 2:5, 1:4, 1:3, 2:2, 3:3, 4:4 (Brown et al. 1993). The actual translation relationships may be simpler, but noise from repeated instances of closed class words interferes with them. Furthermore, unlike sentence alignment, the relationships in slot alignment are often many-to-many and the slots that should be aligned often do not come in the same order

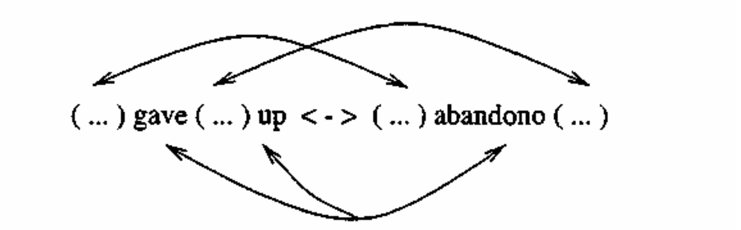

Figure 2: Representation of the task of string and slot alignment

3.3.1 Aligning Slots and Strings

To align the strings, we use a technique closely related to, but much simpler than that employed by Gale & Church (1993) for aligning sentences. Gale & Church (1993) discovered that the lengths of corresponding paragraphs in bilingual corpora, at least between European languages, were highly correlated. This revealed the fact that length and relative order are good clues in determining translations. From their experiments, they deduced that longer sentences in one language tend to be translated by longer sentences in another language and ditto for shorter sentences. We have further deduced that, as one moves down the hierarchy of textual units, translation clues based on the length of the textual unit may still apply, albeit not as reliably.

We align the slots and strings by measuring their lengths in characters and finding the closest match between them, based on the ratio between their lengths. In this way, a SL string is aligned with a TL string with the most similar character length. Ditto for slots. In the current state of the algorithm, the position of the slots has no bearing on the alignment. Therefore, straightforward alignments, such as a SL active clause translated by a TL active clause, as in (1), and inverted alignments, as in (10) where a passive clause in English is translated by an active clause in Spanish, are equally likely to be correct.

(10) *(the system) was developed by (a team of engineers) ⇔ (un equipo de ingenieros) desarrollaron (el sistema)*

In this example, the strings in brackets denote the strings that form the slots. In this case, the algorithm correctly aligns *the system* with *el sistema* and *a team of engineers* with *un equipo de ingenieros,* since they have similar character lengths.

This method of alignment is partially successful, given that as the length of the textual unit that we want to align decreases, so does the possibility of its successful alignment. Sentences can be difficult to align, but refining the textual unit even further incurs larger error rates. However, we have used it as a heuristic method and found that it does give useful results, with the simpler slot relationships (1:1, 2:2) being easier to align. The greater the disparity between the number of slots on either language side of the pair and the greater the number of slots to align, the more difficult the task of accurate and meaningful alignment. Disparities often arise from varying structures between languages (such as *N* + *N* compounds being translated as *N* + *prep* + *N* constructions) and noise from repeated occurrences of open class and closed class words in Collocations.

3.3.2 Translation Patterns & Translation Memory

A sentence pattern that has been generalised through place-holders or variables can be more readily identified and used for translation than a whole sentence. It is more flexible than traditional TM in looking for generalised rather than specific sentences and it is also less complex and shallower, requiring less linguistic knowledge than EBMT which relies on parsing source sentences. Consider the following 2 sentence pairs:

Press the Escape key to continue ⇔ *Appuyez sur la clé d'évasion pour continuer*
Press the Return key to continue ⇔ *Appuyez sur la clé de retour pour continuer*

They can be stored as a pattern in the database as one sentence pair:

Press the X_s *key to continue* ⇔ *Appuyez sur* X_t*pour continuer*

where X_s and X_t need to be in translation correspondence. In this case, X is a variable or place holder (Langé et al. 1997) and in our scheme, this would appear as a slot, with the relationship between the slots made explicit.

4 Evaluation

Table 1 represents the results of applying the above algorithm to a corpus[1] of 3000 sentence pairs (or 0.5 Mb) of English and Spanish. The first table shows how varying the frequency and intersection threshold[2] affects the precision of translation patterns. Similarly, for all thresholds, the second table shows how different patterns of slot a-lignment affects precision. The numbers in parentheses indicate the total number of translation patterns evaluated in each instance. These patterns of slot alignment do not include all patterns discovered by the algorithm.

When evaluating these translation patterns, it was unclear how to define recall, and therefore it has not been evaluated. Precision was defined as the proportion of 'correct' translation patterns over the total number of translation patterns. For each translation pattern, 'correct' was defined as whether the strings on each language side of the translation pattern were translations of each other and that the slots were aligned correctly.

[1] WHO AFI corpus at http://wyw.who.int/pll/cat/cat_resources.html

[2] A threshold of 1 indicates the Complements as they represent items that appear possibly once only in the corpus

Threshold	Precision
1 - 4	53% (211/400)
5	76% (76/100)

Slot Alignment Pattern	Precision
1:1	84% (185/220)
2:2	52% (76/146)
1:2	21% (15/72)
2:1	35% (9/26)
0:1	20% (2/10)
1:0	0% (0/22)
0:2	0% (0/2)
2:0	0% (0/2)

Table 1: Evaluation of Translation Patterns

Similarity Score	Quantity
0 - 24	21%
25 - 49	32%
50 - 75	25%
75 - 100	22%

Score	Translation String
0	de emergencia al humanitaria asistencia
25	asistencia técnica para emergencias
50	humanitaria de emergencia a
75	acción humanitaria de emergencia
100	asistencia humanitaria de emergencia

Table 2: Translation Quality

As this judgement is liable to subjectivity, we gave 250 examples of translation patterns selected at random to 5 different native Spanish speakers fluent in English.

From the results, it can be seen that the precision of the translation patterns increases substantially with a threshold of 5. Also, the most easily alignable pattern of slots to match was a 1:1 relationship. Slot alignment patterns of a *X:0* or *0:X* relationship (where $X \geq 1$) frequently failed. The algorithm also has difficulty in matching compounds or other similar divergent structures between languages, as outlined in subsection 3.3.1.

A TM system that makes use of such translation patterns requires a composition step, similar to that in EBMT. Given a SL input, relevant translation patterns combine, using the string and slot alignments, to build TL translations. Table 2 presents preliminary results for such a composition algorithm, the details of which are to be reported in a future publication. A test corpus (10%) of our data was retained, while translation patterns from the remaining 90% were extracted. Each SL sentence from the 10% test corpus was subsequently translated using the patterns extracted from the remaining 90%. Each translation produced was compared with the reference TL sentence as defined by the test corpus. The process was repeated for the remaining nine ways of selecting a test corpus. It was possible to produce translations for 83% of the 3000 test sentences, including partial or fuzzy matches. The results in the first table are percentages of this 83%.

A dynamic programming (DP) algorithm was used to compare each translation (TR) produced by the algorithm with the reference translation (RT) from the test corpus. Mis-matches of words, insertions and deletions were each penalised with a score of one. Using the formula below, a similarity score was calculated for each comparison, normalising against the length of the longest of the RT or the TR, as this represents the greatest dissimilarity score possible. A value between 0 and 100 is obtained as a similarity score or percentage of similarity. Given a SL sentence, the highest similarity score from the set of comparisons is returned as the result for that translation. The second table provides an indication of the quality of translations produced for a given set of similarity scores. The translation of *emergency humanitarian assistance* is presented. A 100% similarity score represents a perfect match with the TL sentence in the corpus.

$$\left(1-\frac{DP(RT,TR)}{Longest(RT,TR)}\right)100$$

5 Future Directions

First, in an effort to increase the rate of correct slot and string alignments in translation patterns, lexical information from ‘safe’ previous alignments can be used to build a bilingual lexicon in order to improve alignments. Such a bilingual lexicon could also be constructed using a method such as that in Dagan et al. (1993) as this is 'knowledge-free'. Second, these experiments have been carried out without using linguistic knowledge. We plan to add a small amount of knowledge specific to a language pair in the form of stop lists and include stemming and word similarity metrics. Finally, improving the composition stage is our next major area of research.

6 Acknowledgements

This research is funded by the UK Economic and Social Research Council, award number R00429834560.

References

Ahrenberg, L, M. Andersson & M. Merkel: 1998, ‘A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts’, in *Proceedings of the 17th International Conference on Computational Linguistics: COLING-98,* Montreal, Canada, pp. 29-35.

Benson, M: 1990, ‘Collocations in General-Purpose Dictionaries’, in *International Journal of Lexicography* 3: pp. 23-35.

Brown, P.F, S.A. Della Pietra, V.J. Della Pietra & R.L. Mercer: 1993, ‘The Mathematics of Statistical Machine Translation: Parameter Estimation’, in *Computational Linguistics,* 19: pp. 263-311.

Brown, R.D: 1997, ‘Automated Dictionary Extraction for “Knowledge-Free” Example-Based Translation’, in *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation,* Santa Fe, New Mexico, July 1997. pp. 111 - 118.

Chen, M.H, J.S. Chang & J-N Chen: 1997, '*Top-Align:* Word Alignment for Bilingual Corpora Based on Topical Clusters of Dictionary Entries and Translations', in *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation,* Santa Fe, New Mexico, pp. 127-133.

Dagan, I, K.W. Church & W.A. Gale: 1993, 'Robust Bilingual Word Alignment for Machine Aided Translation', in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives,* Columbus, Ohio, pp. 1-8.

Fung, Pascale & McKeown, Kathleen: 1997, 'A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups', in *Machine Translation,* 12: pp. 53-87.

Gale, W.A & K. W. Church: 1993, 'A Program for Aligning Sentences in Bilingual Corpora', in *Computational Linguistics,* 19: pp. 75-102.

Gaussier, Eric: 1998, 'Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora', in *Proceedings of the 17th International Conference on Computational Linguistics: COLING-98,* Montreal, Canada, pp. 444-450.

Kaji, H., Y. Kida & Y. Morimoto: 1992, 'Learning Translation Templates from Bilingual Text', in *Proceedings of the 15th International Conference on Computational Linguistics: COLING-92,* Nantes, France, pp. 672-678.

Langé, J-M, E. Gaussier & B. Daille: 1997, 'Bricks and Skeletons: Some Ideas for the Near Future of MAHT', in *Machine Translation* 12: pp. 39-51.

Simard, M, G. F. Foster & P. Isabelle: 1992, 'Using Cognates to Align Sentences in Bilingual Corpora', in *Quatrième colloque international sur les aspects théoriques et méthodologiques de la traduction automatique, 4th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-92,* Montréal, Canada, pp. 67-81.

Smadja, Frank: 1993, 'Retrieving Collocations from Text: Xtract', in *Computational Linguistics,* 19: pp. 143-177.

Smadja, F, K. McKeown & V. Hatzivassiloglou: 1996, 'Translating Collocations for Bilingual Lexicons: A Statistical Approach', in *Computational Linguistics* 22: pp. 1-38.

Somers, Harold: 1998, 'Further Experiments in Bilingual Text Alignment', in *International Journal of Corpus Linguistics* 3: pp. 115-150.

Takeda, Koichi: 1996, 'Pattern-Based Context-Free Grammars for Machine Translation', in *Proceedings of the 34th Meeting of the Association for Computational Linguistics* Santa Cruz, California, pp. 144-151.

van der Eijk, Pim: 1993, 'Automating the Acquisition of Bilingual Terminology', in *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics* Utrecht, The Netherlands, pp. 113-119.

Watanabe, Hideo: 1993, 'A Method for Extracting Translation Patterns from Translation Examples', in *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-93: MT in the Next Generation,* Kyoto, Japan, pp. 292-301.