32. Preprocessing multilingual corpora

1. Introduction

A multilingual parallel corpus is a corpus that contains the same text samples in each of at least two languages, in the sense that the samples are translations of one another. A comparable corpus contains texts in at least two languages which are not translations of each other, but are written in the same genre and on the same topic (see article 16). A sentence alignment is a mapping showing which sentence or sentences of the one text correspond with which sentence or sentences of the other language. Word-level alignment is also possible, and valuable for the production of terminological dictionaries. The focus of this article will be on the preprocessing of parallel corpora, and in particular, on the preprocessing steps of tokenisation (identifying the constituent words, sentences and paragraphs of a corpus, see article 24) and automatic alignment. Figure 32.1, actually a modified segment of the Canadian Hansards adapted from Simard et al. (2000), shows two types of translation analyses − horizontal lines to denote the segmentations of a sentence alignment and numbers in parentheses to denote a word-level mapping. The Canadian Hansards are transcripts of the proceedings of the Canadian Parliament, recorded in both English and French (cf. article 20).

7(1).	7(1).
Le développement(2) est maintenant appréhendé dans la multiplicité de ses dimensions(3).	Development(2) is now understood to involve many dimensions(3);
On n'y voit plus seulement un problème de politique économique(4) et de resources.	it is no longer merely a matter of economic(4) politics and resources.
Les facteurs politiques, sociaux, éducationnels et environmentaux sont concus comme autant de facettes de l'action unifée(5) à développement.	Political, social, educational and environmental factors must be part of an integrated(5) approach to development.
Au moins(6) d'un développement a l'échelle la plus vaste, les jeunes seront agités, frustrés et improductifs.	Without(6) development on the widest scale, the young will be restless, resentful and unproductive.
On se disputera les resources(7) et la creéativité s'égarera.	People will fight for resources(7) and creativity will be misdirected.

Fig. 32.1: A short section of a bilingual corpus, aligned at the sentence and word-level.

2. Applications of aligned corpora

Aligned corpora are useful for foreign language learners. They can be used in conjunction with concordancers to show language learners real examples of how a construction in one language has been translated on various occasions into the other (Barlow 2000). They are also useful for bilingual terminology extraction. Multinational organisations such as the European Union continually need to translate product documentation and standardise terminology in technical fields. Human acquisition of technology is an expert task, both slow and expensive, and it is difficult to keep up with the pace of technology development. The production of printed dictionaries involves an inevitable time lag, and commercial dictionaries do not typically contain the subject-specific vocabulary which is often needed. The automatic derivation of bilingual terminology lists offers a solution to both these problems (Gaussier/Langé 1994). Machine Translation, which works best in highly specific technical domains such as weather forecasting (Kittredge 1985), is an important area where such subject-specific electronic dictionaries are particularly useful.

Parallel corpora have a number of other applications in machine translation. With example-based machine translation (EBMT), a large parallel corpus of previously translated phrases is stored (cf. article 56). When a new phrase is entered, the most similar phrase in the corpus is found, and its translation returned (Nagao 1984). Transfer rules, which determine how syntactic structures in one language are translated into the syntactic structures of the other language in traditional machine translation, can also be derived from parallel corpora (Carbonell et al. 2002). Parallel corpora form the basis of statistically-based machine translation (Brown et al. 1990). Parallel corpora can also act as the "gold standard" against which the output of machine translation systems is compared. They also provide data for theoretical linguistic studies (as discussed in article 54).

3. Sentence segmentation and word tokenisation

Most alignment algorithms for parallel corpora require the preprocessing steps of tokenisation (identification of word boundaries) and sentence segmentation (clearly marking where one sentence ends and another begins). Obviously there can be no alignment at the sentence level if we do not know where all the sentence boundaries are, and if we do not identify all the word boundaries correctly, our lexicons derived from word-level alignment will contain a number of non-word sequences (Palmer 2000). In reality, there is no absolute definition for what constitutes a word or a sentence. For example, should *seaside* or *sea side* be one word or two? A lively account of this from a linguist's point of view is given by Aitchison (1999). There are differences between the ways different corpora and applications segment words and sentences. For example, the British National Corpus (BNC) regards *President's* as two word tokens, *President* and *'s*, while the Brown corpus would regard it as a single word token (cf. article 24).

Gale/Church (1991) use manual marking up of sentences (denoted *-d*) and paragraphs (denoted *-D*) before their sentence alignment program can begin. Sentence segmentation involves the recognition of boundaries, typically punctuation (such as a full stop or question mark) at the end of a sentence. Making this assumption, most errors will arise by confusion between full stops denoting abbreviations and those denoting sentence

boundaries. There are various heuristics (none absolutely fail-safe) which use the context surrounding the full stop to decide whether it is more likely to denote an abbreviation or end of sentence, such as, does the next word start with an upper case letter? Full stops can also show decimal points (when they are surrounded by numeric characters), and salutations such as *Mr.* Raynar/Ratnaparkhi (1997) suggest that titles such as *Mr.* do not occur at the end of a sentence, and certain suffixes suggest that words are not likely to be abbreviations. Palmer/Hearst (1997) found that the sequences of the parts-of-speech (POS) of the words surrounding the full stop give clues as to its function. Speech quote marks also cloud the issue of sentence segmentation.

Kiss/Strunk (2002) also describe a method for determining whether a full stop denotes an end of sentence or an abbreviation. They regard a word followed by a full stop as a collocation of that word and the full stop. The strength of this collocation, as measured by the log likelihood measure, is greater if the word is an abbreviation, because that abbreviation will always be followed by a full stop. Other words will be weakly collocated with the full stop, since they often occur at other places in a sentence, and thus are not always followed by a full stop. This method will not distinguish between an abbreviation at the end of a sentence and an abbreviation elsewhere in the sentence.

Two successful alignment algorithms circumvent the need for word and sentence segmentation. One is the K-vec algorithm of Fung/Church (1994), where the corpus is simply divided into segments of equal length. The other is Char_align (Church 1993) where alignment is at the level of sequences of four adjacent characters, which may or may not span word boundaries.

4. Sentence alignment

Many algorithms for bilingual sentence alignment exist, and Wu's (2000) excellent chapter on alignment extracts some important general principles. The task of sentence alignment is to discover exactly which sentence or sentences in the first language correspond to which sentence or sentences in the other language. It is often the case that there is no 1:1 alignment between sentences in two languages. For example, Figure 32.2 shows an alignment between two English sentences and one French sentence (i. e. a 2:1 alignment). Strictly speaking, an alignment must be monotonic, meaning that coupled passages (referred to as beads) must occur in the same order on both sides of the corpus. Monotonic sentence aligners will generally propose spurious alignments during non-monotonic passages in translations. Sentence order is generally preserved in translation, but this is not the case for word order. In this article we will use the phrase "word-level alignment", but since word order is generally not preserved in translation, strictly speaking we should say that the words are "set in correspondence".

4.1. Constraints for alignment

The imposition of constraints such as monotonicity helps us a great deal, because myriads of putative alignments are at a stroke declared impossible, cutting down the number of possible alignments from which we must choose the best. Initially there are so many

possibilities that even the computer could not look at them all in a reasonable amount of time. For example, to align just 20 sentences of one language with 20 sentences of another, assuming that only 1:1 alignments were allowed (but allowing crossover), would mean having to try about 2,430,000,000,000,000,000 possible configurations before deciding on the best.

Other constraints are anchors**,** or points in each text which are known to correspond with each other. Some anchors are given by the document structure, and thus tend to be corpus specific, such as labels to identify dates and speakers in the Canadian Hansards, and numbered section headings. It is generally assumed that the very beginning and very end of the two texts are good anchor points. Other anchors are bilingual lexical constraints. In a few cases we can be certain that a word in one language is always translated as the same word in another language, and thus the two sides of the corpus must align wherever this word pair is found. Although less directly useful, we can also make use of word pairs which are mutual translations most, but not all, of the time. Section 5 on word-level alignment describes how such pairs can be discovered and assigned numeric scores according to their translation reliability. Heuristics also exist for the identification of cognates (see section 4.4.) which can also act as partially reliable anchors for related language pairs such as English and French. A related constraint heuristic is referred to as "bands"**,** where it is assumed that "the correct couplings lie not too far from the diagonal between adjacent anchors" (Wu 2000).

One group of alignment methods, described as lexical methods (Kay/Röscheisen 1993; Catizone/Russell/Warwick 1989), make much use of anchor points, especially bilingual word correspondences. They have the advantage that they are more robust to noisy texts (i. e. tolerant of imperfect translations); however, the most easily implemented and fastest corpus alignment techniques are based on relative sentence lengths, which use only paragraph boundaries as anchor points.

4.2. Alignment algorithms based on relative sentence lengths

Alignment methods based on sentence length (Brown et al. 1991; Gale/Church 1991, 1993) operate on a very simple, but surprisingly reliable, premise: short sentences in one language tend to be translated by short sentences in another language, and long sentences in one language tend to be translated by long sentences in the other. Thus the greater the difference between the length of a sentence in language A and the length of a sentence in sentence B, the less likely that they should be aligned. To take this into account, a numerical penalty or cost is imposed whenever the algorithm considers aligning two text segments which differ in length, as given in the following formula:

$$\delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}}$$

l_1 and l_2 are the lengths, in characters, of the two segments. c is a factor which takes into account that equivalent texts might on average be longer in one language than the other. This must be determined empirically for each language pair for which the algorithm is to be used, and can be found simply by dividing the number of characters in

the corpus for language A by the number of characters for language B. Gale/Church (1993) found this ratio was 1.06 when language A was French and language B was English. s^2 is the variance in the number of characters in language B per character in language A. If there were always exactly 1.06 characters of French for every character of English, this variance would be 0. However, if we sometimes find other ratios of lengths in characters in individual translation sentence pairs, the variance will be more than 0 (in fact it is 5.6). As we will see in section 8, the variance is high for English and Chinese.

The next stage is to calculate Prob(δ), which is the proportion of sentences which are translations of each other with δ degree of length difference or more. Many statistics textbooks have a table for computing this "integration of a standard normal distribution". Alternatively, Gale and Church give the Abramowitz and Stegun approximation which is in a suitable form for inserting into a computer program. The penalty P is taken to be $P = -100 \ln (\text{Prob}(\delta))$ which is capped at 2500 to prevent having to deal with infinitely large values.

There is also a penalty for bead (or block of mutually aligned sentences) type or cardinality − the more rare the bead type proposed, the higher the penalty. The penalty for a 1:1 pairing (substitution) is 0 (i. e. there is no penalty at all), since this is the most common bead type, 2:1 or 1:2 groupings (contractions and expansions) have a penalty of 250, while 0:1 or 1:0 unpaired sentences (insertions and deletions) have a penalty of 450. Finally, merges, where two sentences of English in total translate two sentences of French, but neither of the English sentences exactly translates either of the French sentences, have a penalty of 440. These penalties were derived empirically, using the formula

$$P = -100 \ln \left(\frac{\textit{probability of alignment type}}{\textit{probability of } 1:1 \textit{ alignment}} \right)$$

The two penalties (for sentence length and cardinality) are added together to give the overall cost of a single bead. The overall cost of aligning the whole bitext is the sum of the costs of the constituent beads, as shown in the example of Figure 32.2. Gale and Church's program could be extended in principle to include any n:m bead type.

Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1998.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minerals, entrée en vigeur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.
Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	

Fig. 32.2: A short section of the Canadian Hansards

There are three possible alignments of this corpus. Firstly, we might have a 2:1 alignment, where the combination of the two English sentences is translated by the single French sentence. The English sentences are 124 and 106 characters long, while the length of the French sentence is 267 characters respectively. First we consider the scenario where the two English sentences are translated by the single French sentence. The combined length of the English sentences is 230 characters. Using the formula given by Gale/

Church (1993), the penalty for this small difference in text lengths is calculated as 57. The penalty for a 2:1 coupling is 250, so the total cost of this alignment is the sum of the two penalties, 57 + 250 = 307.

Secondly, we might have a 1:1 substitution followed by a 0:1 insertion, i.e. the French sentence is completely translated by the first English sentence only, while the second English sentence has no French translation. The overall cost of this alignment is the sum of the costs of the two constituent beads, i.e. 880 + 2404 = 3284. Thirdly we might have a 0:1 insertion followed by a 1:1 substitution, where the first English sentence has no French translation, and the second English sentence alone fully translates the French sentence. The overall cost of this is 3868. These values show that the first scenario, a 1:2 expansion, is the likeliest of the three, since it has the lowest overall cost. For the sake of completeness, we will mention that there is a fourth, albeit unlikely scenario, where none of the text portions are translations of each other. Here the corpus would be divided into three beads, one for the deletion of first English sentence, one for the deletion of the second English sentence, and one for the insertion of the French sentence (various orderings of these are possible, with no difference in cost). The cost of this would be prohibitive, at 8063.

In this small example we have only four possible alignment combinations from which to select the best. In a real corpus, there would be simply too many possible alignment configurations for them all to be examined in turn in a reasonable time, and thus a technique called dynamic programming is used. This means that only certain alignments deemed to be reasonably likely from the outset are examined using the Gale and Church formulas. Theoretically the optimal solution might be overlooked, but a good solution will be found within a reasonable length of time.

4.3. Dynamic programming for sentence alignment

If we wish to align e sentences of English with f sentences of French by dynamic programming, we should first produce a matrix with $(e + 1)$ rows and $(f + 1)$ columns. The rows are numbered from 0 to e, and the columns from 0 to f. We start with the trivial assumption that the cost of aligning no sentences of English with no sentences of French is 0, and store this value in row 0, column 0, of the matrix. In every case, the value stored in row x, column y, is the minimum cost of aligning x sentences of English with y sentences of French. There is only one way of aligning a sequence of sentences of English with no sentences of French, and that is by a sequence of deletions. Thus the value in row 1, column 0, is the cost of deleting the first English sentence. We add to this value the cost of deleting the second English sentence, and put this result in row 2, column 0. In this way we can easily fill up all the squares in column 0. As well as recording the cost of these partial alignments, we also keep a record that the last step taken in each case was a deletion. Analogous reasoning allows us to fill up all the squares in row 0, by working out the costs of sequences of insertions.

Now we come to fill in the inner cells of the matrix. The value in row 1, column 1, should be the minimum cost of aligning the first sentence of the English text with the first sentence of the French. The word "minimum" now becomes important, because we have a choice of three ways of achieving this alignment: a single substitution, a deletion

followed by an insertion, or an insertion followed by a deletion. The cost of a single substitution can be measured directly. To find the cost of a deletion followed by an insertion, remember that we have already found the cost of the deletion, which is stored in row 1, column 0 of the matrix. To this we simply add the cost of inserting the first French sentence. Similarly, to find the cost of an insertion followed by a deletion, find the cost of the insertion at row 0, column 1, then add the cost of the deletion. These three values are compared, and the smallest is retained as the value in row 1, column 1. Once again, as in every case, we record the nature of the most recent bead (substitution, insertion or deletion) to be added to the developing alignment.

Deeper inside the matrix, we have to consider all 6 bead types. In Figure 32.3, the cost of aligning the first three sentences of English with the first two sentences of French is derived from the costs of 6 earlier alignments plus the cost of adding one more bead. The least of these 6 values is the cost of the alignment.

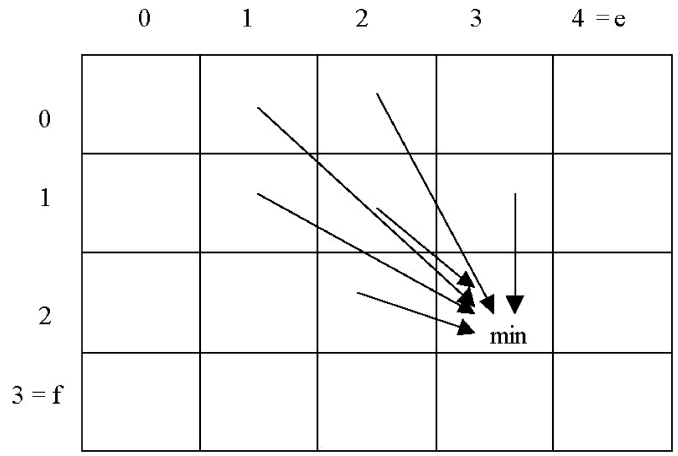

Fig. 32.3: Dynamic programming matrix for sentence alignment

Tab. 32.1: Calculating the cost of a 3 : 2 alignment

Old align	Min cost	Bead to make up 3 : 2	Cost of new bead	Total cost of 3 : 2
1 : 0	450	2 : 2	440	890
2 : 0	900	1 : 2	250	1150
1 : 1	0	2 : 1	250	250
2 : 1	250	1 : 1	0	250
3 : 1	700	0 : 1	450	1150
2 : 2	0	1 : 0	450	450

Table 32.1 shows a set of possible values for the situation described in Figure 32.3. These values are calculated using the Gale and Church penalties for the various bead types, but to simplify the example, penalties based on sentence lengths are ignored. The first column shows the six alignments at the tails of the arrows in Figure 32.3, which can all

be transformed into 3 : 2 alignment by the addition of just one more bead. In the second column, the previously found minimum costs of achieving each of these alignments are given. For example, the least cost of producing a 2 : 2 alignment is by proposing two consecutive 1:1 alignments, with a total cost of 0. The third column shows the bead type required to transform each alignment in the first column into a 3 : 2 alignment, and the fourth column shows the cost of adding this particular bead type. The final column shows the overall cost of producing a 3 : 2 alignment by each of the six possible routes (rows in the table). In fact we have a tie, with either a 1 : 1 bead followed by a 2 : 1 bead, or a 2 : 1 bead followed by a 1 : 1 bead, with least overall cost (250). Usually such a tie would be resolved by considering the relative sentence lengths, but if not, whichever of the lowest cost alignments is considered first by the program will be chosen.

In general, to find the cost of a partial alignment of x sentences of English with y sentences of French, where the last bead type to be added was of type $a : b$ (coupling a sentences of English with b sentences of French), proceed as follows: Go back to row $(e-a)$, column $(f-b)$, to find the minimum cost of the previous sequence of beads, then add the cost of adding the $a : b$ bead type. It is necessary to work through the matrix systematically, such as from left to right, top to bottom. When all the cells have been filled, the value in row e, column f is the cost of the optimal alignment. We have also kept an implicit record of the sequence of bead types which produced this optimal alignment, and we can retrieve this as follows: See which bead type was the last to be added to the final alignment. If this was an $a : b$ bead type, we trace back to row $(e-a)$, column $(f-b)$, to find the penultimate bead type. This process is repeated until we arrive back at row 0, cell 0.

To overcome the problem of the dynamic programming technique sometimes finding a non-optimal alignment, Chen/Chen (1994) used a technique from artificial intelligence called simulated annealing. The idea of annealing comes from physics, where in order to render some material such as enamel into its most stable configuration, it is first heated, and then slowly cooled again until it falls into this configuration. Heat can be represented in an alignment program by randomness, and the most stable configuration by the optimal alignment, which has least energy (corresponding to least overall cost). Chen/Chen start with any possible alignment, such as one previously found by dynamic programming, which may well not be the best. A number of moves are allowed in the search for a better alignment, e. g. take a sentence from one bead and move it into an adjacent bead. The quality of each alignment is the number of matching part-of-speech (POS) tags on each side. The likelihood of a putative new alignment being adopted is greater if it is better than the old, but it will not always be accepted. Similarly, the randomness in the system means that it is possible for a poorer alignment to be adopted. The reason that seeming improvements are not always made, and even some poorer alignments are chosen, is to prevent the system getting stuck in what is called a local minimum − a sort of dead end, which might be the best among a set of alignments just one step removed from each other, but poorer than the overall best which may be completely different. The search for the optimal alignment continues in this stepwise manner. The degree of randomness in the system gradually gets less until "better" steps are always made, and "worse" steps are never made.

Dynamic programming is also used in bioinformatics for the alignment of protein and DNA sequences (Kruskal 1983), and may even be used by the human body itself for protein folding (Hockenmaier et al. 2006).

4.4. Incorporation of cognates into sentence alignment

Simard/Foster/Isabelle (1992) used the term cognates to describe a pair of words, each from a different language, which are orthographically similar and have similar meaning. This definition would include similar words in languages which are not historically related, such as borrowings and proper nouns, although a linguist's definition of cognates would require historical relatedness. Simard et al. were the first to suggest that the discovery of such word pairs could assist in the process of sentence alignment.

McEnery/Oakes (1996) suggested that such cognates could be discovered in corpora using approximate string matching techniques, such as Dice's similarity coefficient (Dice 1945) and the Damerau-Levenshtein metric (Damerau 1964). Dice's similarity coefficient was adapted by Adamson/Boreham (1974) so it could be used as a measure of orthographic similarity between two words. McEnery/Oakes (1996) used their technique to describe the degree of similarity between a word in one language and its translation in another, such as the Italian *avverbio* and the English *adverb.* First the two words are separated into lists of their adjacent character pairs, or bigrams, as follows: *av-vv-ve-er-rb-bi-io* and *ad-dv-ve-er-rb.*

The number of matching bigrams (*ve*, *er* and *rb*) is 3, while the total number of bigrams in both words is 5 + 7 = 12. Dice's similarity coefficient is twice the number of matches, divided by the total number of bigrams in the two words − which is 6/12 = 0.5. The formula returns a value of 1 if the two words are orthographically identical, and 0 if they have no character pairs at all in common. Empirical results for English-French (McEnery/Oakes 1996) and English-Polish (Lewandowska-Tomaszczyk/Oakes/Wynne 1999) show that the greater the similarity coefficient between a word in one language in a bilingual corpus and a word in the other, the more likely it is that these two words are mutual translations. Hofland (1996) used a combination of bilingual word pairs found by Dice's similarity coefficient with those found in a Norwegian-English dictionary to guide sentence alignment.

Similarly, cognates can be identified using the Damerau-Levenshtein metric, which is the smallest number of operations required to transform one word form into another. This technique is a variant of the Gale and Church sentence alignment algorithm in microcosm. Once again dynamic programming is used, but now we are performing alignment at the character level. Only three operations are allowed: insertions, deletions and substitutions. Thus three operations (a substitution of $v \rightarrow d$ and deletions of *i* and *o*) are required to transform *avverbio* into *adverb*. The number of operations can be transformed to a 0 to 1 scale by dividing by the length in characters of the longer word − in this case two words identical in form will have a score of 0, while two totally dissimilar words will have a score of 1.

Simard et al. (2000) discovered the importance of using only isolated cognates to guide the process of sentence alignment, where a word form is isolated if no similar word form occurs within a certain number of characters. The identification of non-isolated cognates will degrade alignment performance rather than improve it. Cognates which appear in the same aligned region of a corpus (and hence the same context) are unlikely to be 'false friends'. Other linguistic clues can be incorporated into alignment algorithms, such as part-of-speech information and the results of shallow syntactic analysis. Piperdis et al. (1994) showed that words are more likely to be aligned if they take the same part of speech, and Tiedemann (2003) gives the example that a verb phrase in English is more likely to match a verb cluster in Swedish than a noun phrase.

4.5. Dealing with noisy bitexts

A number of techniques have been designed for the alignment of noisy bitexts (bilingual parallel corpora) or poor quality translations. Wu (2000) lists some reasons why these might occur: non-literal translation, out of order translation, omitted sections, floating passages (such as footnotes, figures and headers), optical character reader (OCR) errors, and sentence segmentation errors. The key to the success of these techniques is that they do not require the original texts to be accurately broken up into paragraphs. The output is a set of anchors rather than a complete mapping of sentences. One such technique is the char_align program of Church (1993), which aligns texts at the character level, using the dot plot technique. Texts should be anchored at character *x* in English and character *y* in French if the tetragrams (sequences of four characters) leading up to characters *x* and *y* are identical. If so, a dot is placed in cell (*x,y*) of the dotplot. This only works if there are sufficient cognates in the two languages being aligned. Simard found that about 21 % of the words in a 100 sentence sample of mutual translations in the English-French Canadian Hansards are cognates, which is sufficient for the dotplot technique to work. In contrast, only about 6 % of the words were cognate in a sample of about 100 randomly chosen sentence pairs which were not mutual translations.

Areas which should not be aligned are sparsely populated with dots (due to random correspondences), while a more pronounced diagonal line appears to show where the true correspondences should lie. These diagonal lines can be enhanced by image processing techniques (Chang/Chen 1997). Melamed (1997) describes alignment of noisy bitexts by combining scores for both statistical and linguistic comparisons, such as the presence of cognates, bilingual dictionary matches, and part of speech information.

4.6. Evaluation techniques for sentence alignment

Sentence alignment, like sentence segmentation, can be evaluated using Recall and Precision (Simard et al. 2000). The machine-aligned corpus is compared with a humanly-aligned reference corpus. Precision is the number of sentence pairs aligned by both human and machine, divided by the number of sentence pairs aligned by the machine only. Recall is the number of sentence pairs aligned by both human and machine, divided by the number of sentence pairs aligned by the human only. Here a 2 : 1 alignment of sentences of 10 and 11 in English with sentence 10 in French would mean that the sentence pairs (10, 10) and (11, 10) would be aligned. Both Recall and Precision can also be used with the number of characters aligned by each technique, so the results are not unduly influenced by very short sentences. Gale/Church (1993) stress the importance of the human judge marking the alignments beforehand, so as not to be influenced by seeing the machine output first. This also allows comparison of results for different algorithms with different output formats on a common basis.

5. Word alignment

The two main reasons for performing alignment at the word level are a) to help guide a process of sentence alignment, b) for building bilingual lexicons automatically, and c) to train statistical machine translation systems on the word-aligned corpora.

The earliest work on word alignment was done by Brown et al. (1990), where it was achieved as a by-product of statistical machine translation. Word alignments were produced by the multiplication of three sets of automatically discovered probabilities: a) translation probabilities, e. g. the English "not" aligns with the French "pas" with probability 0.469; b) fertilities, e. g. "not" corresponds with two words in French with probability 0.758; and c) distortion, e. g. what is the probability of the second word in an English sentence aligning with the fifth word of a French sentence?

A number of sentence alignment processes, such as that of Kay/Rőscheisen (1993) work in a series of iterated steps. First a rough estimate of which sentences might be aligned is made, which provides enough information to make a first pass at word alignment. Having found a small number of word pairs which are reliable translations, it is possible to make an improved pass at sentence alignment, which in turn enables better word-level alignment. This process continues until no new word pairs can be found.

Daille et al. (1994) built bilingual terminology banks from the International Telecommunications Corpus. The sentence-level alignment of Figure 32.1 was derived first, and this enabled the later word-level alignment. Daille (1995) lists a wide range of statistical measures, all based on the contingency table shown in Figure 32.4. N = a + b + c + d, which is the total number of aligned regions in the corpus (6 in the example of Figure 32.1).

a: number of aligned regions in which both the French and the English words appear	*b*: number of aligned regions in which the French word appears but the English word does not appear.
c: number of aligned regions in which the English word appears but the French word does not appear.	*d*: number of aligned regions in which neither the French nor the English word appears.

Fig. 32.4: A contingency table for word alignment

All these measures are based on the principle that if an English word is the translation of a French word (or vice versa), then in most cases, wherever we have the French word on the left hand side of the corpus, we also have the English word on the right hand side, in the same aligned region. Thus high values of *a* give strong evidence that two words are probably translations of each other, while high values of *d* give weak evidence that they are translations. High values of *b* and *c* (meaning many aligned regions where only one of the two words is found, not both) are evidence that the two words are not translations.

The simplest measure is simple co-occurrence frequency, value *a* in the contingency table. If we want to discover automatically the most likely translation of the English word *development,* it must be one of the words found in a French sentence aligned with at least one English sentence containing *development.* Examples of such words are *le, développement, est, maintenant, apprehendé* and *dans.* Since *développement* and *development* co-occur 3 times (more often than any other French word co-occurs with *development*), the simple co-occurrence frequency suggests that *développement* is the most likely translation of *development*. More sophisticated measures exist, to take into account that the very common words (such as *la, de*, and *les* in this example) tend to co-occur frequently with every English word, simply because they are so common and are found

throughout the corpus. These more complex measures take not only *a* into account, but also the other values of the contingency table. Daille adapted the simple mutual information measure to produce the cubic association coefficient MI^3, which is given by the formula

$$MI^3 = \log_2 \left(\frac{a^3 N}{(a + b)(a + c)} \right)$$

In the small corpus of Figure 32.2, the MI^3 between *development* and some of its more promising translation candidates in the French part of the corpus is as follows: *développement* 4.17, *la* 2.41, *les* 2.41. Other candidate words, such as *maintenant*, have a maximum MI^3 of 1. Thus the MI^3 measure, like simple co-occurrence frequency, suggests that *développement* is the most likely translation of *development*. Using this method, the most likely translation can be found for each word in the corpus, to build up a complete lexicon. The most effective measures, as determined empirically by Daille, were simple co-occurrence frequency, MI^3, the Fager/McGowan (1963) coeffiecient, and the Log Likelihood measure (Dunning 1993). Fung/Church (1994) suggest a double check: only selected word pairs which have both high MI and t-scores.

Nowadays, researchers wishing to use a "ready-made" word aligner often use the GIZA++ system of Och/Ney (2000).

5.1. Enhancements to statistical word alignment

A number of enhancements have been made to the general procedure described above. Gaussier/Langé/Meunier (1992) were able to eliminate some incorrect high scoring pairs using their best match criterion. For example, they originally created a list of possible candidate translations of the English word *prime* in the Canadian Hansards, consisting of French words found to have a high mutual information score (MI, similar to MI^3, but with the *a* on the top line not cubed) with *prime.* At the top of the list were *sein* (5.63), *bureau* (5.63), *trudeau* (5.34), and *premier* (5.25). Using their best match criterion, they eliminated all candidates which had been found to have a higher MI score with an English word other than *prime*. This left the word *premier* (the correct translation) at the top of the list.

Word-level alignment as described here can be enhanced by stemming rules, for the removal and replacement of common prefixes and suffixes, designed to render alternative grammatical forms of a word equivalent. In the corpus segment shown in Figure 32.2, we have two related pairs of words occurring just once each, *politique* in the same aligned region as *politics*, and *politiques* in the same aligned region as *political.* Since these word pairs occur only once each, we have only weak evidence that they represent translation pairs ($MI^3 = 2.58$). However, if we recognise that in fact we have two matching pairs of words derived from *polit-,* we now have stronger evidence that these form a translation pair ($MI^3 = 3.58$). Thus the use of stemming rules can help us better identify translation pairs. French stemming rules have been developed by Savoy (1993), and the most commonly used set of English stemming rules are those of Porter (1980). Similar improvements can be achieved by lemmatisation, where each word is reduced to its dictionary

headword. In the calculations presented here, a word has been taken to be any sequence of characters surrounded by white space. Better results would be obtained by preprocessing the corpus to split word pairs forming contractions such as *s'impose* → *se impose*, *l'échelle* → *la échelle*.

When deriving word-level alignments from corpora already aligned at the sentence level, Gaussier/Langé/Meunier (1992) found about 65% of English words were assigned their correct French translations, 25% had no French word assigned (mainly words with no real French equivalent) and about 10% were aligned with words that were not their correct French translations.

If a bilingual corpus (where the two texts are translations of each other) has not been aligned at the sentence level, it is still possible to use the statistical measures described in this section for word-level alignment. Fung/Church's (1994) K-vec method requires only that the corpus be cut into sections of equal length, and corresponding sections be treated as aligned regions for word length alignment. More complex, and less direct statistical measures can be used for comparable corpora, which contain texts in two languages on the same topic, although they are not translations of each other (Fung/Yee 1988; Peters/Picci 1998; Gaussier et al. 2004). Here we must start with a small bilingual dictionary, which is augmented with a new word pair from the corpora whenever it can be shown that sufficient collocates of this new word pair are found to correspond in the dictionary.

5.2. Matrix factorisation techniques for word alignment

Goutte/Yamada/Gaussier (2004) achieve word-level alignment from a corpus already aligned at the sentence level, by aligning the words of the two languages through central pivots called cepts, which roughly correspond to individual concepts. More than one word can be aligned with a single cept. The sentences *the licence fee does not increase* and *le droit de permis ne augmente pas* align via four cepts as follows*: the (1) le; licence fee (2) droit de permis; not (3) ne pas; increase (4) augmente*. One matrix is created to store the alignments between the English words and the cepts, and another for the mappings between the cepts and the French words.

These two can be multiplied together to produce the translation matrix, which stores the strength of association between each English word and each French word. An earlier word alignment technique which depended on matrix multiplication was given by Tanaka/Iwasaki (1996).

6. Phrase alignment

Gaussier/Langé (1994) also used statistical measures to work on the problem of finding correspondences between technical terms which were collocations of two content words (such as *station terrienne* and *earth station*) in an aligned International Telecommunications Union (ITU) corpus. The MI between this pair of technical terms was taken to be sum of the following: the MI (as derived above) between *station* and *earth*; the MI

between *station* and *station;* the MI between *terrienne* and *earth*; and the MI between *terrienne* and *station*. This was called mutual information with double association.

Lee/Chang/Jang (2006) were interested in aligning named entities such as the names of people and organisations in bilingual documents, as part of machine translation. English named entities were identified either automatically or manually as a preprocessing step, after which the following steps were repeated until all the English named entities were aligned: a) find the set of translation candidates occurring in the target (Chinese) sentence using phrase translation and transliteration (see below); b) evaluate the set of translation candidates, and sort by translation score; and c) align the source and target named entity pair with highest probability.

Following Brown et al. (1990), the phrase translation phase was decomposed into a lexical translation score (the probability of an individual English-Chinese word pair being translations of each other), and a position alignment score such as P(1 = 2, 2 = 1, 3 = 3), the collective probability of the first English word matching the second Chinese word, the second English word matching the first Chinese word, and the third English word matching the third Chinese word. Following Gale/Church (1993), penalties were given for insertions and deletions (such as if a three-word Chinese named entity was proposed as the translation of a two-word English named entity). Names of people and places are typically transliterated into their phonetic equivalents. Thus the system also learns the rules for transliterating English characters into their Pin Yin (Romanised spelling) equivalents. A bilingual dictionary and parallel corpora were used as training data to obtain both the phrase translation and transliteration probabilities.

Other researchers who have used statistical measures to find correspondences between multi-word units are Haruno/Ikehara/Yamazaki (1996), Kitamura/Matsumoto (1996), Smadja/McKeown/Hatzivassiloglou (1996) and McEnery et al. (1997).

7. Structure and tree alignment heuristics

Alignment is not only possible between linear sequences, such as sentences of linear text, but between tree structures. This is important when aligning parse trees from translated texts, to extract phrases for example-based machine translation. The output of tree alignment is a mapping between pairs of coupled nodes. One heuristic for doing this is the crossing constraint (Wu 2000): Suppose two nodes in language-1 (p1 and p2) correspond to two nodes in language-2 (q1 and q2) respectively, and p1 dominates p2. Then q1 must dominate q2. Matsumoto/Ishimoto/Utsuro (1993) give further heuristics for mapping nodes. A cost is associated with any of their heuristics which are not completely fulfilled, and the alignment with least overall cost is the one chosen.

a) Couple leaf nodes (words) that are lexical translations, as found in a bilingual lexicon.
b) Couple leaf nodes that are similar. For example if we have a node for *cat* in English, the correspondence *cat* → *chat* in the English-French dictionary, and a term related to *chat* in the French thesaurus is *tigre*, then the nodes for *cat* and *tigre* will match.
c) Couple internal nodes that share as many coupled leaf nodes as possible.
d) Couple nodes that share as many coupled children or descendents as possible.

Another early method, that of Grishman (1994), assumes that we have a bilingual corpus which has already been aligned at the sentence level, and that both source and target texts have been independently parsed. We also need a bilingual dictionary which lists typical translations for many (but not necessarily all) of the words in the corpus. Node Si in the source parse tree can only be paired with node Ti in the target parse tree if at least one of the following conditions holds true:

1. Ti is a possible translation of Si as found in the bilingual dictionary.
2. There is at least one pair <Sj,Tj> in the alignment such that Si dominates Sj and Ti dominates Tj.
3. There is a pair in the alignment <Sj,Tj> such that Sj immediately dominates Si, Tj immediately dominates Ti, and the syntactic role taken by Ti is a possible translation of the role taken by Si.

Even when these rules are applied, there will be multiple possible alignments. To choose between them, a score for each overall alignment is assigned, which is the sum of the scores of the individual pairings that make up the alignment, which depend on the following four criteria:

1. If Ti is a possible translation of Sj;
2. Whether Si dominates any other nodes in the alignment;
3. If Si immediately dominates other nodes in S which correspond to nodes in T;
4. For each node Tj in the alignment which is immediately dominated by Ti, is the syntactic role a possible translation of the role filled by the node Sj with which it is paired?

All possible alignments between the two trees are considered, and the highest scoring one is chosen as being the most likely.

The various steps for source tree to target tree transformation allowed in Gildea's (2004) model for tree-tree alignment are:

1. Reordering a node's children;
2. Inserting and deleting nodes;
3. Translating individual words at leaf nodes;
4. A single node in the source tree may become two nodes in the target tree and vice versa.

To find the most likely transformation sequence from among a number of possible alternatives, certain probabilities must be learned beforehand from corpora. Related to transformations (1) and (2) above are the reordering probabilities which are in the form $Palign(\{(1,1)(2,3)(3,2)\} \mid A \rightarrow XYZ)$ meaning "given that the children of node A in the source language appear in the order X, Y, Z, what is the probability of the words corresponding to Y and Z being inverted in the target language?". Such reorderings can include insertions and deletions of individual children. The likelihood of type (3) transformations is given by word to word translation probabilities, and for type (4) we consider the probability of the current node being grouped with one of its child nodes, given the nature of the production rule which decomposes that parent into its child nodes.

In the parse-parse-match scenario, the two sides of the bitext are independently parsed into their constituent structures. Wu (1995), on the other hand, used a biparsing grammar to parse both sides of the bitext simultaneously.

8. Alignment with genetically unrelated language pairs

One of the main challenges when working with languages other than English is character-set dependence. The original set was the ASCII 7-bit set which can encode 128 characters, adequate for English texts, but email systems which still use it require the users of other European languages to circumvent the lack of diacritics by typing, for example, *grüßen* as *gruessen (German 'to greet')*. There are many larger character sets, such as the 8-bit Latin-1 which covers most Latin-based alphabets with diacritics. Chinese and Japanese require a 2-byte (16 bit) encoding because of their much larger character sets. Examples of character encodings for Chinese are GB and Big5, and TIS 620 for Thai.

The EMILLE project (Baker et al. 2004) has produced monolingual corpora for 14 South Asian languages, and also a parallel corpus of 200,000 words of text in English and its translations in Bengali, Gujurati, Hindi, Punjabi and Urdu. The corpus is marked up in CES-compliant SGML (Baker et al. 1998), which includes sentence and paragraph markers, headings and foreign text (e. g. <s> <p>, <head> and <foreign lang= "eng"> to open sections, and <\s> to close a sentence). Some of the texts were typed in directly using the Unicode word processor Global Writer, and Microsoft Word 2000, running on a Windows 2000 machine, is also Unicode compliant. Other documents were available electronically, such as UK government documents which are available in a wide range of languages, on topics including health, social security and housing. As such, they are rich sources for extracting multilingual term banks in these specific domains. Much material in South Asian languages is encoded in various 8-bit formats. A tool called "Unicodify" was produced by the EMILLE project to convert these various 8-bit encodings into the international standard Unicode, and is available on ⟨http://www.ling.lancs.ac.uk/corplang/emille/default.htm⟩.

A number of languages, most notably Chinese and Japanese, consist of unsegmented character sequences without marked word boundaries, and thus automatic word segmentation is generally required before further processing can take place. In particular, sentence alignment algorithms such as that of Brown/Lai/Mercer (1991) rely on estimating sentence length by the number of words in each sentence, which can only be found by segmentation. Problems with word segmentation for Chinese include the low level of agreement between native speakers as to where word boundaries should be, estimated at 70 % by Sproat et al. (1996). There was a recent international competition for Chinese word segmentation algorithms, called the First International Chinese Word Segmentation Bakeoff (Sproat/Emerson 2003). One of the entrants (Wu 2003) used the two stage process of word recognition followed by disambiguation. Words were recognised by matching against a lexicon of named entities and derivational morphology rules for the recognition of grammatical variants. Wu also derived heuristics for the identification of new words, not in the lexicon. The next task was to disambiguate between alternative sequences of words which have been proposed in stage one. This may be done by finding which of the possible interpretations allows the most meaningful syntactic parse.

The Gale/Church (1993) sentence alignment algorithm was designed to be language pair independent. A ratio factor c is included in one of the formulas to represent the mean ratio of characters in the second language to the number of characters in the first language for a translated text. A second factor, s^2, is included to show the amount of variation in this ratio from sentence to sentence. McEnery/Piao/Xin (2000) found that the lengths of sentences in Chinese are much more poorly correlated with their transla-

tion equivalents in English than a closely-related language pair such as English and French, and thus s^2 should be much higher for English-Chinese sentence alignment. Another problem is that the Gale and Church program assumes that only 6 types of sentence pairings (1:0, 0:1, 1:1, 2:1, 1:2 and 2:2) occur between translation pairs. However, Chen/Chen (1994) found it was necessary to consider other pairings such as 3:1 and 4:1, since it is common to find several short Chinese sentences matching a single long English one. The Gale and Church algorithm can be extended to incorporate these, but substantial amendments to their computer program are required.

Historically-related languages share cognates, so the correlation between their sentence lengths will be better, and also the cognates themselves can be used to guide the alignment. However, even languages which are not genetically related will contain some similar words due to borrowings and proper nouns. Some of the difficulties inherent in aligning English and Chinese named entities can be overcome if the Chinese is transcribed into its Romanised equivalent, Pin Yin (Lee/Chang/Jang 2006). Successful English-Chinese sentence alignment has been achieved by a number of authors (Fung/McKeown 1997; Wang et al. 2002; Piao 2002).

Prior word segmentation is also required for many sentence alignment procedures involving Japanese. Sentence lengths do not always correspond for Japanese and English, since Japanese function words show little correspondence with their English counterparts, and politeness particles are not always translated. Utiyama/Isahara (2003) align Japanese and English newspaper articles by first performing alignment at the paragraph level (finding the pairs of paragraphs which best match in terms of containing corresponding bilingual dictionary pairs), and then doing sentence alignment using dynamic programming.

9. Conclusion

In this article, we have looked closely at three important preprocessing steps for multilingual parallel corpora, namely segmentation at the sentence level, tokenisation at the word level and alignment. A pair of texts may be aligned at the paragraph, sentence, phrase, word or character level. As well as aligning linear texts, we can also align bilingual parse trees. Alignment techniques are mainly statistical, but may also incorporate linguistic information. We have considered how alignment techniques are evaluated, and the special requirements when aligning texts from non-European languages.

10. Literature

Adamson, G. W./Boreham, J. (1974), The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. In: *Information Storage and Retrieval* 10, 253–260.

Aitchison, J. (1999), *Teach Yourself Linguistics*. London: Hodder Arnold.

Baker, J. P./Burnard, L./McEnery, A. M./Wilson, A. (1998), Techniques for the Evaluation of Language Corpora: A Report from the Front. In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain, 135–142.

Baker, P./Hardie, A./McEnery, A./Xiao, R./Bontcheva, K./Cunningham, H./Gaizauskas, R./Hamza, O./Maynard, D./Tablan, V./Ursu, C./Jayaram, B. D./Leisher, M. (2004), Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development. In: *Literary and Linguistic Computing* 19(4), 509−524.

Barlow, M. (2000), Parallel Texts in Language Teaching. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 106−115.

Brown, P. F./Cocke, J./Della Pietra, J./Della Pietra, V./Jelinek, F./Lafferty, J./Mercer, R./Roosin, P. (1990), A Statistical Approach to Machine Translation. In: *Computational Linguistics* 16(2), 79−85.

Brown, P./Lai, J./Mercer, R. (1991), Aligning Sentences in Parallel Corpora. In: *Proceedings of the 29th Annual meeting of the ACL*, Berkeley, CA, 169−176.

Carbonell, J./ Probst, K./Peterson, E./Monson, C./Lavie, A./Brown, R./Levin, L. (2002), Automatic Rule Learning for Resource-limited MT. In: *Proceedings of the Association for Machine Translation in the Americas (AMTA-02)*, Tiburon, CA, 1−10.

Catizone, R./Russell, G./Warwick, S. (1989), Deriving Translation Data from Bilingual Texts. In: *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, MI, 1−6.

Chang, J. S./Chen M. H. (1997), An Alignment Method for Noisy Parallel Corpora Based on Image Processing Techniques. In: *Proceedings of the 35th ACL*, Universidad Nacional de Educacion a Distencia (UNED), Madrid, Spain, 297−304.

Chen, K.-H./Chen, H.-H. (1994), A Part-of-speech-based Alignment Algorithm. In: *Proceedings of COLING'94*, Kyoto, Japan, 166−171.

Church, K. W. (1993), Char_align: A Program for Aligning Parallel Texts at the Character Level. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, 40−47.

Daille, B. (1995), *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*. (UCREL Technical Papers 5.) Department of Linguistics, University of Lancaster.

Daille, B./Gaussier, E./Langé, J.-M. (1994), Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: *Proceedings of COLING'94*, Kyoto, Japan, 515−521.

Damerau, F. J. (1964), A Technique for the Computer Detection and Correction of Spelling Errors. In: *Communications of the ACM* 7. New York: ACM Press, 171−176.

Dice, L. R. (1945), Measures of the Amount of Ecologic Association between Species. In: *Geology* 26, 297−302.

Dunning, E. (1993), Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* 19(1), 61−74.

Fager, E. W./McGowan, J. A. (1963), Zooplankton Species Groups in the North Pacific. In: *Science* 140, 453−560.

Fung, P./Church, K. W. (1994), K-vec: A New Approach for Aligning Parallel Texts. In: *Proceedings of COLING'94*, Kyoto, Japan, 1096−1101.

Fung, P./McKeown, K. (1997), A Technical Word and Term Translation Aid Using Noisy Parallel Corpora Across Language Groups. In: *Machine Translation* 12, 53−87.

Fung, P./Yee, L.-Y. (1988), An Information Retrieval Approach for Translating New Words from Non-parallel, Comparable Texts. In: *Proceedings of COLING/ACL 98*, Montreal, Canada, 414−420.

Gale, W. A./Church, K. W. (1991), A Program for Aligning Sentences in Bilingual Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 177−184.

Gale, W. A./Church, K. W. (1993), A Program for Aligning Sentences in Bilingual Corpora. In: *Computational Linguistics* 19(1), 75−102.

Gaussier, E./Langé, J.-M. (1994), Some Methods for the Extraction of Bilingual Terminology, In: Jones, D. (ed.), *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, 14−16 September 1994, UMIST, Manchester, United Kingdom, 242−247.

Gaussier, E./Langé, J.-M./Meunier, F. (1992), Towards Bilingual Terminology. In: *Proceedings of the Joint ALLC/ACH Conference*. Oxford: Oxford University Press, 121−124.

Gaussier, E./Renders, J.-M./Mateeva, I./Goutte, C./Déjean, H. (2004), A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, July 21st−26th, Barcelona, Spain, 526−533.

Gildea, D. (2004), Dependencies vs. Constituents for Tree-based Alignment. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP04)*, July 25th−26th, Barcelona, Spain, 214−221.

Goutte, C./Yamada, K./Gaussier, E. (2004), Aligning Words using Matrix Factorisation. In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL),* July 21st−26th, Barcelona, Spain, 502−509.

Grishman, R. (1994), Iterative Alignment of Syntactic Structures for a Bilingual Corpus. In: *Proceedings of the 2nd Annual Workshop for Very Large Corpora*, Tokyo, Japan, 57−68.

Haruno, M./Ikehara, S./Yamazaki, T. (1996), High Performance Bilingual Text Alignment Using Statistical and Dictionary Information. In: *Proceedings of COLING'96*, Copenhagen, Denmark, 525−530.

Hockenmaier, J./Joshi, A. K./Dell, K. A. (2006), Protein Folding and Chart Parsing. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 22nd, Sydney, Australia, 293−300.

Hofland, K. (1996), A Program for Aligning English and Norwegian Sentences. In: Hockey, S./Ide, N./Perissinotto, G. (eds.), *Research in Humanities Computing.* Oxford: Oxford University Press, 165−178.

Kay, M./Röscheisen, M. (1993), Text-translation Alignment. In: *Computational Linguistics* 19(1), 121−142.

Kiss, T./Strunk, J. (2002), Scaled Log-likelihood Ratios for the Detection of Abbreviations in Text Corpora. In: *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Aug 24th−Sept 1st, Taipei, Taiwan, 1228−1232.

Kitamura, M./Matsumoto, Y. (1996), Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In: Ejerhed, E./Dagan, I. (eds.), *Proceedings of the 4th Workshop on Very Large Corpora*, Copenhagen, Denmark, 79−87.

Kittredge, R. I. (1985), The Significance of Sublanguage for Automatic Translation. In: Nirenburg, S. (ed.), *Machine Translation: Theoretical and Methodological Issues.* Cambridge: Cambridge University Press, 59−67.

Kruskal, J. B. (1983), An Overview of Sequence Comparison. In: Sankoff, D./Kruskal, J. B. (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Reading MA: Addison-Wesley, 1−44.

Lee, C.-J./Chang, J. S./Jang, J.-S. R. (2006), Alignment of Bilingual Named Entities in Parallel Corpora Using Statistical Models and Multiple Knowledge Sources. In: *ACM Transactions on Asian Language and Information Processing*. New York: ACM Press, 121−145.

Lewandowska-Tomaszczyk, B./Oakes, M. P./Wynne, M. (1999), Automatic Alignment of Polish and English Texts. In: Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.), *PALC'99: Practical Applications in Language Corpora*. Frankfurt a. M.: Peter Lang, 77−86.

Matsumoto, Y./Ishimoto, H./Utsuro, T. (1993), Structural Matching of Parallel Texts. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, 23−30.

McEnery, A. M./Langé, J.-M./Oakes, M. P./Véronis, J. (1997), The Exploitation of Multilingual Annotated Corpora for Term Extraction. In: Garside, R./Leech, G./McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman, 220−230.

McEnery, A. M./Oakes, M. P. (1996), Sentence and Word Alignment in the CRATER Project. In: Thomas, J./Short, M. (eds.), *Using Corpora for Language Research.* London: Longman, 211−231.

McEnery, A. M./Piao, S./Xin, X. (2000), Parallel Alignment in English and Chinese. In: Botley, S. P./ McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 177−189.

Melamed, I. D. (1997), A Portable Algorithm for Mapping Bitext Correspondences. In: *Proceedings of the 35th Annual Meeting of the ACL / 8th Conference of the European Chapter of the Association of Computational Linguistics*, Madrid, Spain, 305−312.

Nagao, M. (1984), A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In: Elithorn, A./Banerji, R. (eds.), *Artificial and Human Intelligence*. Amsterdam: Elsevier Science Publishers, 173−180.

Och, F. J./Ney, H. (2000), Improved Statistical Alignment Models. In: *Proceedings of the Association for Computational Linguistics (ACL)*, Hong Kong, 440−447.

Palmer, D. (2000), Tokenisation and Sentence Segmentation. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 11−35.

Palmer, D./Hearst, M. A. (1997), Adaptive Multilingual Sentence Boundary Disambiguation. In: *Computational Linguistics* 23(2), 241−267.

Peters, C./Picci, L. (1998), CLIR: A System for Comparable Corpus Querying. In: Grefenstette, G. (ed.), *Cross-language Information Retrieval*. Norwell, MA: Kluwer, 81−92.

Piao, S. S. (2002), Word Alignment in English-Chinese Parallel Corpora. In: *Literary and Linguistic Computing* 17(2), 207−230.

Piperidis, H./Cranias, L./Papageorgiou, S. (1994), A New Approach to Automatic Sentence Alignment. Poster presented at Teaching and Language Corpora (TALC94), 10−13 April 1994, Lancaster, United Kingdom. Abstract available at: http://www.comp.lancs.ac.uk/ucrel/talc_handbook.ps.

Porter, M. F. (1980), An Algorithm for Suffix Stripping. In: *Program* 14, 130−137.

Reynar, J. C./Ratnaparkhi, A. (1997), A Maximum Entropy Approach to Identifying Sentence Boundaries. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington D.C., 16−19.

Savoy, J. (1993), Stemming of French Words Based on Grammatical Categories. In: *JASIS* 44(1), 1−9.

Simard, M./Foster, G./Hannan, M.-L./Macklovitch, E./Plamondon, P. (2000), Bilingual Text Alignment: Where Do we Draw the Line? In: Botley, S./McEnery, A./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 38−64.

Simard, M./Foster, G./Isabelle, P. (1992), Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI92)*, Montreal, Canada, 67−81.

Smadja, F./McKeown, K./Hatzivassiloglou, V. (1996), Translating Collocations for Bilingual Lexicons: A Statistical Approach. In: *Computational Linguistics* 22(1), 1−38.

Sproat, R./Emerson, T. (2003), The First International Chinese Word Segmentation Bakeoff. In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing,* July 2003, Sapporo, Japan. Available at: http://www.sighan.org/bakeoff2003/paper.pdf.

Sproat, R./Shih, C./Gale, W./Chang, N. (1996), A Stochastic Finite-state Word Segmentation Algorithm for Chinese. In: *Computational Linguistics* 22(3), 377−404.

Tanaka, K./Iwasaki, H. (1996), Extraction of Lexical Translations from Non-aligned Corpora. In: *Proceedings of COLING'96*, 580−585.

Tiedemann, J. (2003), Combining Clues for Word Alignment. In: *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL'03)*, April 12th−17th, Budapest, Hungary, 339−346.

Utiyama, M./Isahara, H. (2003), Reliable Measures for Aligning Japanese-English News Articles and Sentences. In: *Proceedings of ACL'03*, July 7th−9th, Sapporo, Japan, 72−79.

Wang, W./Zhao, M./Huang, J. X./Huang, C. N. (2002), Structure Alignment Using Bilingual Chunking. In: *Proceedings of COLING'02*, Aug 24th−Sept 1st, Taipei, Taiwan, 1−7.

Wu, A. (2003), Chinese Word Segmentation in MSR-NLP. In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing,* Sapporo, Japan, July 2003, 172−175.

Wu, D. (1995), An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words. In: *Proceedings of the 33rd ACL*, MIT, Cambridge, Mass., 244−251.

Wu, D. (2000), Alignment. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 415−458.

Michael P. Oakes, Sunderland (UK)