Evaluating Word Alignment Systems

Magnus Merkel & Lars Ahrenberg
NLPLAB, Department of Computer and Information Science
Linköping University
S-581 83 Linköping, Sweden,
magme@ida.liu.se, lah@ida.liu.se,

1. Introduction

In the last five to ten years the interest in systems that align (or link) words and phrases in a source text to corresponding target units has increased steadily. In this paper we use the notion of *word alignment systems* as a general term for systems that align linguistic units below the sentence level across two languages. These linguistic units could be expressed as single words, phrases, terms or collocations. The majority of the word alignment systems described in the literature fall into two main categories: (1) Full-text alignment systems, and (2) Bilingual lexicon extraction systems. Below these main categories, it is possible to make further divisions into, for example, bilingual concordancing and bilingual information retrieval for the first category, and technical terminology systems and systems that compile lexicons automatically for specific systems or specific uses for the second category.

To be able to compare algorithms and systems that perform word alignment on parallel corpora is fundamental to progress in the area. There are several problematic issues for the evaluation of word alignment systems, the most important being,

- **The purpose of the alignment system**. A program designed for bilingual lexicon extraction differs from a program that aims at aligning a whole text with its translation. Furthermore if the output data is used for bilingual concordance browsing, the system should be evaluated with this aim in mind.
- **Units**. What characterises the translation units? Should multi-word units be counted as such? Should function words be included or excluded?
- **Resources used**. When systems are compared, information on how long it takes to run the system on a particular bitext should be included, as well as extra resources such as bilingual lexicons and monolingual collocation lists.
- **The use of a Gold standard**. When alignment output is evaluated it can be compared to a Gold standard, which is constructed *before* the actual alignment, or experts can evaluate a sample of the output *after* the alignment.
- **Metrics and scoring method**. What metrics should be used? When the output is evaluated, there are several questions on how to judge partial alignments when collocations are involved, deletions, insertions, segmentation errors and paraphrases.
- **Error analysis**. What is the nature of the mistakes that a particular system makes? Does it typically fail on certain types of collocations, on units within a particular frequency range, etc?

In the rest of this paper, we will try to address these issues in relation to word alignment systems in general, but also to full-text alignment systems and lexicon extraction systems specifically.

2. The purpose of the alignment system

Sometimes it is difficult to distinguish between different types of systems which all share the general objective of identifying correspondences between text units in a source and a target text. However, a program that extracts a bilingual lexicon is primarily aimed at finding translations for content units, that is, terms, phrases and content words. On the other hand, one can say that a program that aims at aligning all tokens in a text can also produce a bilingual lexicon. The resulting bilingual lexicon (which is just a generalisation of all the link tokens) will typically contain entries that are not aimed for in a pure lexicon extraction program. The evaluation method should therefore be tailored to a specific type of alignment system in order to avoid unfair comparisons.

3. Units

In a pure word-to-word model (cf. Melamed 1995), many valid lexical units are missed due to the fact that they belong to collocations or complex paraphrases. For all kinds of word alignment linking, it is necessary to be able to handle multi-word segments in both the source and target text. Some approaches use pre-processing on only the source side (Melamed 1997b, Smadja et al. 1997) and then the target correspondences are estimated during the linking stage. In other approaches, both the source and target texts are pre-processed independently and candidate lists for both source and target multi-word units are created to be used in the linking process (cf. Ahrenberg et al. 1998).

The major difficulty is to identify all collocations present in a text, especially when the frequency is low. Furthermore, it is not obvious how to make the segmentation for certain multi-word units, such as particle verbs and prepositional objects.

Recall is also difficult to measure when multi-word units are considered, due to the fact that it is more or less impossible to know how many collocations there are in a text. Recall measurements can therefore only be made on samples of a bitext.

4. Resources used

Some word alignment systems make use of extra resources, such as bilingual dictionaries, function word lists, morphological components, taggers, phrase lists or different separate programs for processing multi-word units. The resources used by a particular system is valid information in the evaluation. Even if a "black box" approach is adopted, and the output is judged against checked reference data, the types of resources a system can utilize are of necessary if a complete picture of the system's performance is to be painted. Information on how long it takes to run the system on a particular bitext is also relevant for the evaluation as well as what platform and hardware that is used.

4. Gold standards

Gold standards are usually a sample of the bitext that has been prelinked manually by one or several annotators and then used to test the alignment output automatically. They come in two main formats:

1. Complete alignment of the sample. This is a method where the source and target sentences in the sample are broken down into segments and the translation correspondences are marked. Melamed (1998) used this method in the Blinker project.
2. "The translation spotting" method. Here a number of word or phrase types from the source text are chosen. All the sentence pairs that contain the singled-out tokens are presented to the annotator who chooses the corresponding target word/phrase. This method was used in the evaluation phase of the Word track part in the Arcade project during the summer of 1998 (Veronis 1998).

The advantage of the first method is that nothing can be avoided. All the text segments in the sample have to be annotated. The disadvantage is that it can be hard to arrive at a single correct mark-up, especially if you have several annotators. Melamed reports that the inter-annotator agreement was 82 per cent if function words were included in the sample and 92 per cent if only content words were considered. A great deal of work therefore has to be put into creating unambiguous instructions that guide the annotators. Furthermore, the technical problems involved in the mark-up of a gold standard could be considerable. Often a dedicated system has to be created from scratch that checks that everything that should be annotated also is done. And then the result from the annotators' work should be compiled into one single Gold standard to which the output from the alignment system can be checked easily and accurately.

The second method, the Arcade style, makes it possible to cover different types of words and phrases in a more consistent way. In the Arcade competition, 60 word types were singled out, 20 verbs, 20 adjectives and 20 nouns. Here all word types had a frequency of around sixty, but they were chosen on the basis that they exhibited some kind of interesting problem concerning polysemy. By varying the selection criteria, the translation spotting method could help to evaluate units over various dimensions, such as frequency ranges, polysemy and parts of speech.

5. General metrics and scoring methods

The standard metrics used for measuring the performance of NLP systems is recall and precision. A proposed alignment *A* of a bitext can be measured against a reference alignment A_r (for example a Gold standard). The recall of the alignment A with respect to the reference alignment A_r can be defines as:

$$recall = \frac{|A \cap A_r|}{|A_r|}$$

The precision of the alignment is then defined as follows:

$$precision = \frac{|A \cap A_r|}{|A|}$$

The precision measurement gives the proportion of segments in the proposed alignment *A* that is considered to be correct.

The above recall and precision measurements are straightforward to handle if the text only consists of single words, but it becomes increasingly more difficult when the alignments are not one-to-one, which they indeed are not when collocations are involved, as well as for deletions, insertions, segmentation errors and paraphrases.

The scoring for precision and recall can be adjusted to handle partial alignments by using some kind of weighted scores.

6. Evaluation of Full-text alignment

When the output of an alignment system is some kind of encoding of all lexical units and their corresponding translations, there are basically two ways to evaluate the output:

1. A posteriori evaluation of a sample of the output.
2. Comparison with a Gold standard that has been constructed in advance.

Measurements like recall and precision should be calculated, but there are known problems with this, which involves partially correct links and overlapping of segments.

Token linking and the evaluation of such linking could be limited to certain types of words; for example the evaluation could be restricted to only content words.

In previous approaches to full-text alignments several methods of scoring have been used. One approach is to measure the results relative to a gold standard, using either a sample of continuous text (Melamed (1998) or spot checks (Veronis 1998) (see section 4 for further details). Another way is to evaluate the type links created by the full-text alignment system as a bilingual dictionary and measure recall and precision based on a sample of this dictionary (cf. Ahrenberg et al. 1998). A drawback of such a method is that it is difficult to judge non-standard correspondences by the evaluators when the word pair is presented in a dictionary without its context. Furthermore, a type link may appear correct in the dictionary, even though it is based on erroneous link instances in the text.

One way used by Kitamura & Matsumoto (1996) is to regard precision only on the highest ranked *n* hundred candidates suggested by the system. Recall was then measured relative to the set of words that occurred at least twice in the corpus (i.e., the candidates above the frequency threshold built into the system). Gaussier (1998) uses a similar approach where the top 500 links are checked for precision. Here only precision figures are given; recall is not possible to measure as the set of candidates is not exhaustive. Both Kitamura & Matsumoto and Gaussier had the links checked after the alignment was done. The differences in the way precision and recall is used by various researchers are also illustrated by, for example, Kaji & Aizono (1996) who define recall as the proportion of all word correspondences in the bilingual corpus that are actually extracted. In some approaches, partial links are included in the overall score for precision and in some they are not.

In the Word alignment track of the ARCADE project (Veronis 1998), the scoring methods used for evaluation try to address exactly these problems. Each link (whether it consists of a single word or a multi word unit) is evaluated by calculating precision and recall in the following way.

Precision=Correctly proposed words/Proposed words
Recall=Correctly proposed words/Total no. reference words

In Table 1 below an illustration is given for three link instances. The first unit in the reference, “a/b/c” (i.e. a three-word phrase) is then compared with what the system proposed, namely “a/d”. This gives a precision of 0.5 and a recall of 0.33 for this particular instance. If the system fails to propose a candidate when there exists an actual translation, both the precision and recall scoring will be zero. If a certain unit is not translated and the system also fails to find a link, the precision and recall figures will be 1 for such instances. The total precision and recall rates are then calculated as the average of all the link instances in the sample.

Table 1 Example of precision and recall scoring in ARCADE

Reference words	**Proposed words**	**Precision**	**Recall**
a/b/c	a/d	1/2=0.5	1/3=0.33
e/f	-	0	0
-	-	1	1
Average		0.5	0.44

The advantage of the above scoring method is that the successful linking of multiword units is visible and rewarded, but also that partially correct linking is not deemed out entirely. The disadvantage of this approach is that the reference words and the proposed words are only considered from a the point of view of the target. It presupposes that the source units in the reference are the same as the ones that the system has tried to link with corresponding target units. To remedy this drawback, a comparison of the source segmentation between the reference and the alignment system should be performed.

If evaluations of full-text alignment systems are to be really useful, it is not sufficient to know how well they perform in terms of precision and recall. An evaluation should also contain information on what the strengths and weaknesses of the particular system are. Therefore a predefined set of categories would help to describe the characteristics of the alignment. Perhaps it would be possible to tie this set to the different dimensions that can be tested with a gold standard of the Arcade style.

7. Evaluation of bilingual lexicon extraction

If the purpose is to evaluate an extracted lexicon, i.e. a set of link types, there are several ways to do this. For example:

1. Compare the type links to an existing bilingual lexicon.
2. Have the output evaluated by (lexicographical) *experts* posteriori.
3. Have the output evaluated by *laymen* posteriori.
4. Measure the “explanatory power” of the extracted lexicon by applying it on an already defined sample of the corpus (Melamed 1997a)

5. Limit the evaluation to certain kinds of type links based on the purpose of the evaluation, for example terminology, content words, etc.

In some work presented in the literature, the explicit goal has been to extract bilingual dictionaries (e.g. Klavans & Tzoukermann 1990 and Fung & McKeown 1996). Bilingual terminology extraction can also be seen as a kind of specialization of bilingual extraction systems (e.g. Dagan & Church 1994). As mentioned earlier in section 1.1, several methods can be used to evaluate an automatically created dictionary, for example by automatic comparison with a machine-readable bilingual dictionary or by having the extracted dictionary evaluated by experts. However, a comparison of an extracted dictionary with an existing bilingual dictionary could give the wrong results. Non-standard translations, translations of collocations, technical terminology, etc. are often not found in standard dictionaries which, as a consequence, will produce misguiding scoring measurements.

As far as we can tell, there are no standard way of calculating scores in terms of precision and recall for extracted bilingual lexicons, but an alternative evaluation method could be a more pragmatic and practical approach, similar to the solutions suggested by Dagan & Church (1994) and Fung & McKeown (1996). Both adopt a way to measure the increase in efficiency that can be observed when translators are using a particular machine-extracted dictionary. The translators who tested the dictionaries extracted by Fung and McKeown, for example, increased the number of correct term translations by 47 per cent.

As an extension in the same vein, i.e. a practical kind of an evaluation, one could imagine a scenario where professional lexicographers use automatically extracted dictionaries to update commercial bilingual dictionaries. The lexicographical database that contains the commercial dictionary would be compared with the extracted dictionary and suggestions of possible new entries for the database would be presented to the lexicographers who in turn can choose whether or not the new entry should be added. By comparing how many entries that are actually added by using such a technique with the “old” way of updating dictionaries would prove a valuable evaluation of automatically extracted dictionaries.

The same information about strengths and weaknesses as mentioned for full-text alignment systems applies to lexicon extraction systems. Problems that arises from, for example, segmentation and stemming could be included in this set of pre-defined categories for lexicon extraction systems.

8. Conclusions

There is considerable interest in text alignment on the word and phrase level, but some confusion on how alignment systems should be evaluated. The first, and perhaps most important, step is to decide the purpose and usage of such a system. If it is to be adopted for creating full-text alignments used for bilingual searches (bilingual concordancing) or for creating bilingual dictionaries, the evaluation must be tailored towards that particular usage. Secondly, the appropriate segmentation of the source text, in particular, is fundamental for comparisons of scorings between different systems. The metrics used for evaluating systems is often varying between different approaches, even for systems with same overall goal. One solution, at least for full-text alignment systems, may be the use of gold standards, where a correct reference is set up and against which the system output is measured. Further information, in addition to the scoring results, is also of interest and should be included in evaluations. This includes information on the type of errors that the system makes and also information about system performance in terms of time and memory usage as well as data on the implementation and hardware. In this paper several approaches to evaluation of alignment systems have been described with regard to the purpose of the system, text segmentation, metrics and scoring methods, gold standards, error analysis and performance data.

References

Ahrenberg, L, M. Andersson & M. Merkel (1998) A simple hybrid aligner for generating lexical Correspondences in Parallel Texts. In Proceedings of COLING-ACL-98, Montreal, pp 29-35.

Dagan, I, and K W. Church (1994) "Termight: Identifying and Translating Technical Terminology." In *Proceedings from the Conference on Applied Natural Language Processing (ANLP-94)*, 34-40.

Fung, P. & K. McKeown (1996) *A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups*. The Machine Translation Journal, Special Issue on New Tools for Human Translators, pp. 53-87.

Gaussier, E. (1998) Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *Proceedings of COLING-ACL-98*, Montreal, pp. 444-450.

Kaji H. & T. Aizono (1996) Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. In *Proceeding of COLING-96*, Copenhagen, pp. 23-28.

Kitamura, M. & Y. Matsumoto (1996) Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In *Proceedings of WVLC-96*, Copenhagen, pp. 79-87.

Klavans J. & E. Tzoukermann (1990) Linking Bilingual Corpora and Machine Readable Dictionaries with the BICORD System. In *Proceedings from the 6th Annual Conference of the UW Centre for the NEW OE Dictionary and text research*, pp. 19-30.

Melamed, I. D. (1997a) A Word-to-Word Model of Translational Equivalence. In *Proceedings of ACL-97*, Madrid.

Melamed, I. D. (1997b) Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, Providence, RI.

Melamed, I. D. (1998) *Manual Annotation of Translational Equivalence: The Blinker Project*, IRCS Technical Report #98-07, 1998.

Smadja, F., K. McKeown, & V. Hatzivassiloglou (1996) "Translating Collocations for Bilingual Lexicons: A Statistical Approach." In Computational Linguistics, Vol 22. No. 1.

Veronis, J. (1998) *ARCADE - Evaluation of parallel text alignment systems* URL: http://www.lpl.univ-aix.fr/projects/arcade/index-en.html.

Sparck-Jones, K. & J. R. Galliers (1995) *Evaluating Natural Language Processing Systems*. Springer, Berlin.