Evaluation of Word Alignment Systems

Lars Ahrenberg*, Magnus Merkel*,
Anna Sågvall Hein†, Jörg Tiedemann†

* Department of Computer and Information Science
Linköping University, Institute of Technology,
S-581 83 Linköping, Sweden
{lah, magme}@ida.liu.se

†Department of Linguistics
Uppsala University,
Box 527, S-751 20 Uppsala, Sweden
anna@ling.uu.se, joerg@stp.ling.uu.se

Abstract

Recent years have seen a few serious attempts to develop methods and measures for the evaluation of word alignment systems, notably the Blinker project (Melamed, 1998) and the ARCADE project (Véronis and Langlais, forthcoming). In this paper we discuss different approaches to the problem and report on results from a project where two word alignment systems have been evaluated. These results include methods and tools for the generation of reference data and a set of measures for system performance. We note that the selection and sampling of reference data can have a great impact on scoring results.

1. Introduction

Recent years have seen a lot of interest in word alignment systems, i.e., systems that automatically find lexical correspondences in a parallel text (bitext). Such systems have many uses, e.g. in multilingual lexicography and terminology, in contrastive linguistics and translation studies, and in machine translation. There have been some projects and concerted actions to develop methods and measures to evaluate the performance of word alignment systems, notably the Blinker project (Melamed, 1998) and the ARCADE project (Véronis and Langlais, forthcoming), but so far there is very little in terms of common standards and resources for this purpose. This paper should be seen as a contribution to the development of such standards, based on our experience of the ARCADE project and a recent effort to develop tools and methods for the evaluation of our system, the PLUG Word Aligner (PWA).

Word alignment can be viewed as a retrieval problem. The task of the system is to find all correspondences at the lexical level that exist in a given parallel text or corpus. For this reason it seems appropriate to apply measures from the field of information retrieval such as recall and precision to estimate performance. However, the definition of these measures for word alignment systems is not as straightforward as one might expect. Without complementary information, global estimates of recall and precision are likely to be misleading as they depend on properties of the text and the way samples have been generated.

The paper is organised as follows: Section 2 introduces the evaluation problem for word alignment systems and argues for the use of reference data; In Section 3 we discuss three different proposals for measuring recall and precision and review their merits and drawbacks; Section 4 gives a brief overview of our system; In Section 5 we present some data and conclusions from our evaluations of the PWA systems and in Section 6 we state our conclusions.

2. The Evaluation Problem

The most important factors that need to be taken into account for the evaluation of word alignment systems are the following:

- The purpose of the system
- Basic method: prior or posterior reference
- Treatment of multi-word tokens
- Annotation tools and guidelines
- Resources used
- Metrics and scoring methods

2.1. The purpose of the system

There are different types of systems which all share the general objective of identifying correspondences between text units in a source and a target text. 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, a program that aims at aligning all tokens in a text can also produce a bilingual lexicon. For this reason we see performance at the level of link instances as more basic and we focus our discussion on system performance at this level. Obviously, there will be many link instances of function words that are irrelevant for the construction or enlargement of a bilingual lexicon. Thus, evaluation of word alignment systems must allow scoring on different selections of the vocabulary based on frequency, parts of speech and so on.

2.2. Prior or posterior reference

When alignment output is evaluated it can be compared to reference data (usually called a gold standard), which is constructed before the actual alignment, or experts can evaluate the output after the alignment. While gold standards require the creation of tailor-made software, this is not necessary for posterior evaluation. On the other hand, posterior evaluation has the drawback that new data sets must be inspected every time

the system has been run. Thus, in the long run using a gold standard is more efficient and time saving. Also, the criteria used in constructing the gold standard can be determined independently of the design of a given system.

There are, however, many things to consider in the creation of gold standards. The main problem is the compilation of appropriate samples from the bitext that shall be included in the standard. Issues that have to be considered include

- The size of the gold standard
- The distribution of the samples
- The type of sample words

The characteristics of the gold standard are connected to the purpose of the system to be evaluated.

Full-text alignment systems have to include function words whereas systems for bilingual lexicon extraction aim at alignment of content words. This has to be considered in the composition of the gold standard. Furthermore, the sampling method is decisive for the characteristics of the standard. Random samples of reference words produce a high percentage of repeated function words in the gold standard. Usually this induces an extended size of the standard in order to obtain a representative sample. Alternatively, random segments, such as sentences, can be sampled that have to be annotated completely. This might simplify the handling of insertions, deletions, and paraphrasing in the text under consideration.

A lexical evaluation requires a different compilation of reference words in the gold standard. Depending on the task, the standard should include random samples of content words or word instances in a certain frequency range.

2.3. Treatment of multi-word tokens

It is necessary to be able to handle multi-word segments in both the source and target text. One case of multi-word correspondences appears where one language uses syntactic means to express a certain feature and the other language uses morphological means, i.e. a multi-word expression will be linked to a single word. A common case is the English definite article corresponding to a definite suffix in Swedish yielding pairs such as *the car : bilen*, in the pair below:

```
John jumped into □□□□□□□.
John hoppade in i □□□□□.
```

Other common cases of multi-word tokens corresponding to single tokens are compounds such as Eng. ***railway accident*** to Sw. ***tågolycka***, and genitives such as Eng. *the countries* ***of Europe*** to Sw. ***Europas*** *länder.*

Furthermore, correspondences between multi-word tokens in both languages can be found in phrasal constructions (e.g. idioms, fixed expressions and multi-word abbreviations such as '***money talks***', '***after all***', ***'e. g.***') and referring expressions (e.g. proper names and specific terms such as '***New York Times***', ***'ABS control unit***'). It might be also preferred to include further phrasal constructions for specific purposes such as example-based machine translation. This has to be considered in the construction of gold standards as well.

2.4. Annotation Tools and Guidelines

When gold standards are used it is necessary to develop tools for their construction and detailed guidelines for annotation. Both the Blinker and ARCADE projects (Melamed 1998, Véronis and Langlais forthcoming) have made good progress for these tasks. In PLUG a Java-based tool, the PLUG Link Annotator PLA, has been developed for the creation of gold standards (see section 4.2).

2.5. Resources used

Information on how long it takes to run the system on a particular bitext is obviously relevant for the evaluation as well as the platform and hardware that are used. In addition, word alignment systems often make use of extra resources in a preprocessing stage, such as function word lists, bilingual dictionaries or separate programs for identifying multi-word units. The types of resources a system requires have to be considered in the evaluations as they may have a significant influence on the system's performance.

3. Evaluation Metrics

The standard metrics used for measuring the performance of NLP retrieval 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 defined 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 above recall and precision measurements are straightforward to handle if the text only consists of single words, but it becomes more difficult when the alignments are not one-to-one, which they indeed are not when collocations, deletions and additions are involved.

A major difficulty is to identify all multi-word units present in a text, especially those that are discontinuous or have a low frequency; it is more or less impossible to know exactly how many multi-word units there are in a text. Recall measures can therefore in practice only be made on samples of a bitext. Furthermore, multi-word segments give rise to many cases of partial matches that can be treated in different ways in scoring.

Deletions and additions also give problems. Since most available systems are unable to identify them anyhow, one might argue that they should be excluded from reference data. On the other hand it would certainly be valuable if the system can distinguish words that have been translated from those that have not. For this reason, deletions have been considered in the PLUG project (though additions have not). However, there is no way to distinguish a non-response (i.e., when the system has

failed to provide a target link) from when the system determines that there is no link on the target side (e.g. because of a deletion in the translation). We handle this, as proposed in the ARCADE approach (see below), by treating null responses as a special word, "null", which is measured like any other words. From the point of view of the reference, a null link actually represents a deletion, but the current word alignment system cannot provide information about whether it has detected "the deletion" or whether it has just failed to provide a response. This ambiguous feature of null links makes them hard to judge.

3.1. Translation Spotting: P_{ARCADE} and R_{ARCADE}

In the ARCADE word alignment track, the definition of recall and precision were tailored towards the translation spotting problem, e.g. a sub-problem of full text alignment where the system task is restricted to finding the translations of certain source expressions that contain one item from a given list of word tokens. Hence, the reference data and the proposed links are only considered from the point of view of the target. Recall and precision, according to ARCADE, are defined as follows:

$$recall_X = \frac{C_{trg}(X)}{G_{trg}(X)}, precision_x = \frac{C_{trg}(X)}{S_{trg}(X)}$$

$$recall_{ARCADE} = \frac{\sum_{X=1}^{n} recall_X}{n}$$

$$precision_{ARCADE} = \frac{\sum_{X=1}^{n} precision_X}{n}$$

C_{trg} – number of correctly proposed target tokens in link X
S_{trg} – total number of target tokens proposed by the system for link X
G_{trg} – total number of target tokens in the gold standard in link X

Consider the examples in table 1 for an illustration of the scoring method in ARCADE.

Reference words	Proposed words	Precision	Recall
general safety rules	safety booklet	1/2=0.5	1/3=0.33
fault codes	-	0	0
-	-	1	1
Average		0.5	0.44

Table 1. Precision and recall scoring in ARCADE

Translation spotting assumes given source language segments that have to be linked. However, in full-text alignment systems the source text segmentation is not necessarily given in advance. Therefore, a different approach is required, where links rather than target words are counted.

3.2. Full-text alignment: P_{PLUG} and R_{PLUG}

The problem of approximating precision and recall in word alignment systems is to deal with partially correct link proposals and null links properly. Partially correct links include proposals that contain missing parts on the source and/or the target side as well as proposals that go beyond the segmentation borders on the source and/or the target side. We also regard links as partial if they have been proposed "indirectly", i.e., if a multi-word link in the reference has been linked in smaller units and not as one single unit. For example, if the reference states that the command "Save As" should be linked to "Spara som" directly, and the system has suggested the links "Save"-"Spara" and "As"-"som", this link is considered to be indirect, and therefore partial.

The first approach taken within the PLUG project divided the links proposed by the system into four different categories: *incorrect links* (**I**), *correct links* (**C**), *partially correct links* (**P**), and links that were *missed* by the system (**M**). Null links are handled as proposed in the ARCADE approach. Thus, a system 'null' is treated as correct when the gold standard says 'null', otherwise it is regarded as a missing link. Now, the number of links that fit into each of these categories can be counted when comparing the alignment proposed by the system with the appropriate gold standard. Finally, precision and recall can be estimated using these counts. We decided to "penalize" partiality by adding a decreasing weight to the calculation of precision for partially correct alignments. For simplicity the weight was set to 0.5 in our investigations on the PLUG corpus (Ahrenberg et al., 1999):

$$recall_{PLUG} = \frac{n(I) + n(P) + n(C)}{n(I) + n(P) + n(C) + n(M)}$$

$$precision_{PLUG} = \frac{0.5n(P) + n(C)}{n(I) + n(P) + n(C)}$$

n(X) – total number of links in category X

This approach handles partially correct alignment proposals in a very simple way by treating them with a constant score reduction. Although these estimations are quite useful and applicable as shown in Ahrenberg et al. (1999), their limitations can be seen easily. The actual quality of partially correct links is not taken into account. Small mistakes are penalized as hard as big linking errors with small overlapping parts.

	source	target	Q	$\text{precision}_{\text{ARCADE}}$	$\text{recall}_{\text{ARCADE}}$
reference	Reläventil TC	TC relay valve			
proposed	Reläventil TC	Relay valve TC	(3/5 = 0.6) + (2/5 = 0.4) = 1	3/3 ≈ 1	3/3 = 1
reference	ordinarie	ordinary			
proposed	ordinarie skruv	ordinary	2/3 ≈ 0.66	1/1 = 1	1/1 = 1
reference	kommer att indikeras	will be indicated			
proposed	det kommer att indikeras	will the indicated	(2/7 ≈ 0.286) + (0/7 = 0) + (2/7 ≈ 0.286) ≈ 0.572	2/3 ≈ 0.66	2/3 ≈ 0.66
reference	vill	wants			
proposed	-	-	0	0	0
reference	vatten	-			
proposed	-	-	1	1	1
reference	to	till			
proposed	to	att	0	0	0
reference	Scanias chassier	Scania chassis			
proposed	Scanias	Scania chassis	3/4 = 0.75	2/2 = 1	2/2 = 1
$\text{precision}_{\text{PLUG}}$ = (0.5*4+1)/6 = **0.5** **$\text{recall}_{\text{PLUG}}$** = 6/7 ≈ 0.857			**$\text{precision}_{\text{PWA}}$** ≈3.98/6 ≈ **0.663** **$\text{recall}_{\text{PWA}}$** ≈3.98/7 ≈ **0.569**		**$\text{precision}_{\text{ARCADE}}$** ≈4.66/7 ≈ **0.67** **$\text{recall}_{\text{ARCADE}}$** ≈4.66/7 ≈ **0.67**

Table 2. Precision and recall approximation for partial links; some examples.

3.3. A new proposal: P_{PWA} and R_{PWA}

To remedy the coarseness of the PLUG measures we propose the following new metrics:

$$Q = \frac{C_{src} + C_{trg}}{\max(S_{src}, G_{src}) + \max(S_{trg}, G_{trg})}$$

$$recall_{PWA} = \frac{\sum Q}{n(I) + n(P) + n(C) + n(M)}$$

$$precision_{PWA} = \frac{\sum Q}{n(I) + n(P) + n(C)}$$

C_{src} – number of overlapping source tokens in (partially) correct link proposals, C_{src}=0 for incorrect link proposals
C_{trg} – number of overlapping target tokens in (partially) correct link proposals, C_{trg}=0 for incorrect link proposals
S_{src} – number of source tokens proposed by the system
S_{trg} – number of target tokens proposed by the system
G_{src} – number of source tokens in the gold standard
G_{trg} – number of target tokens in the gold standard

Using the definitions above, partially correct alignments are considered proportionally to the number of words that describe the difference between the gold standard and the proposed alignment. Inclusions are similarly included in the precision value as well as links that miss parts compared to the gold standard. Consider the examples in table 2 for a better understanding. Note the ability of these measures to handle partially correct proposals in cases of inclusions as well as in cases of missing parts. The main difference to the approach proposed by ARCADE is the inclusion of the source side in the calculations.

A remaining problem is that alignments that were proposed in terms of sub-links may be twisted (although this case is not very common from our experience). Consider the following link as an example from a gold standard:

▫▫▫▫▫▫▫▫▫▫▫▫▫▫▫ → ▫▫▫▫▫▫▫▫▫▫▫▫▫

Both, $\text{precision}_{\text{PWA}}$ and $\text{recall}_{\text{PWA}}$, would accept sub-links like '***dangerous → gods***' and '***goods → farligt***' to be completely correct even though they are obviously not. This phenomenon is due to the special treatment of linked sub-parts of reference links. A possible solution could be the introduction of decreasing weights as in $\text{precision}_{\text{PLUG}}$ in case of link proposals that do not match the reference link completely. This includes links that do not have the same segmentation as in the gold standard on the source or target side even if the complete reference link is covered by the system's proposals. The problem is to find a proper value for this kind of weight. Empirical investigations might help.

To summarize the pros and cons of the three different metrics disussed in this section the following points could be made:

The ARCADE metrics estimates how successful a word alignment system is for the task of translation spotting by focusing on the relation between the proposed target words with the reference data. However, the ability of the system to correctly segment the source text into single-word and multi-word units is not captured.

The PLUG metrics gives approximate estimations on the system's capabilities to handle partially correct links for full-text word alignment on both the source and target side. Partially correct segmentation and linking are 'penalized' by a standard constant. The scoring will also consider whether a link has been created in one step (direct linking) or in several steps (indirect linking), see section 3.2.

The Q measure attempt to give a more detailed account of a word alignment system for both the source

and target side. However, the Q metrics, like the ARCADE metrics, do not distinguish between direct links and indirect links.

3.4. Global metrics

As will be illustrated below recall and precision calculated over samples of bitexts do not reveal all interesting aspects of system performance. Complementary information is given by measures such as type and token coverage of the source text (the percentage of source types and source tokens that the system has managed to align, whether correctly or incorrectly) and size of the extracted dictionary (number of generated links).

4. PLUG tools and resources

4.1. The PLUG Word Aligner PWA

PWA integrates two systems, the Linköping Word Aligner (LWA; Ahrenberg et al., 1998) and the Uppsala Word Aligner (UWA; Tiedemann, 1998) within the modular corpus toolbox Uplug (Tiedemann, forthcoming). It was developed within a co-operative project on parallel corpora, PLUG (Sågvall Hein, forthcoming).

The objective of PWA is to find link instances in a bitext and to generate a non-probabilistic translation lexicon from them. The system provides output of both kinds.

The system takes input in the form of a bitext divided into segments. Common segments are sentences or sequences of sentences that have been aligned previously for both halves of the bitext. The system combines knowledge-lite approaches to word alignment such that it can be adapted to new language pairs easily. However, information about function words and morphology patterns for each language is required for improving the system's performance.

Furthermore, the system provides modules for the automatic recognition of multi-word units, which is essential for word alignment purposes. PWA handles multi-word correspondences by means of prior generation of word collocations for each language (Merkel & Anderson, forthcoming). It also supports dynamic construction of multi-word units within the linking process (Tiedemann, forthcoming).

The system is iterative, repeating the same process of generating translation pairs from the bitext, and then reducing the bitext by removing the pairs that have been found before the next iteration starts (Melamed 1997, Tiedemann 1997). The algorithm will stop when no more pairs can be generated, or when a given number of iterations have been completed. Links are established by means of co-occurrence measures, string similarity investigations, and other extraction techniques. The approaches are described in more detail in the papers that were mentioned above.

The system is implemented in Perl with versions for Linux, Sun Solaris, and Windows. It can be licenced for academic research purposes from the PLUG home page[1] in a binary version for all the three computer platforms.

[1] The PLUG home page is located at http://stp.ling.uu.se/~corpora/plug/.

4.2. The PLUG Link Annotator PLA

Within PLUG a Java-based tool, the PLUG Link Annotator PLA, has been developed to facilitate the creation of gold standards (Merkel et al., forthcoming). Using PLA human annotators can generate a set of source tokens from a bitext and with the aid of a graphical user interface create reference links. The annotator is free to add any token from the source and target language segment to the link. The annotation is stored in a straightforward format that uses segment identifiers and byte-spans for the identification of each link.

The human annotators create the gold standards with PLA according to a detailed set of guidelines (Merkel, 1999). The two most basic guidelines were the same as those used in the ARCADE project (Véronis and Langlais, forthcoming), namely

- As many tokens as are required to obtain an equivalence should be included in a correspondence;
- No more tokens than are required to obtain an equivalence should be included in a correspondence.

4.3. The PWA Scorer

The PWA system includes an additional module for the automatic evaluation of word alignment processes, the PWA scorer. It applies gold standards as they were discussed above. The alignment software produces a set of link instances from the bitext that includes information about the origin of the aligned units in the text. The scorer reads reference links as they were produced from the PLUG Link Annotator and compares them with the alignments, which were actually found by the system. As the result, the scorer prints an evaluation protocol that includes information about the comparison with each reference link, a summary of the result, and the scores that were calculated by means of different precision and recall metrics. Figure 1 shows a small sample of an evaluation protocol that was generated by the PWA scorer.

```
type      id   source             target
-----------------------------------------------
partial:  77   stärkas            reinforced
                           (be reinforced,1(1)/2)
correct:  147  ansvar             responsibility
incorrect:185  skall (skall)      be (will)
missed:   209  att                to
correct:  229  värld              world
correct:  229  socialt            social
correct:  253  svenska välfärden  swedish welfare
===============================================
number/step      all   2   3   4   5   6
number gold:     100
number returned:  55  10   4   3  36   5
===============================================
number correct:   32   5   4   2  20   1
number null:       2
number partial:   17   4   0   0  13   3
===============================================
number incorrect:  6   1   0   1   3   1
number missed:    43
===============================================
precision (ARCADE):                  49.000%
precision (PLUG):                    74.561%
precision (PWA):                     76.207%
===============================================
recall (ARCADE):                     44.218%
recall (PLUG):                       56.122%
recall (PWA):                        42.284%
===============================================
F-measure (ARCADE):                  46.486%
F-measure (PLUG):                    64.041%
F-measure (PWA):                     54.389%
===============================================
```

Figure 1. A sample of an evaluation protocol.

Beside the final scores, important information about each reference link is provided in such evaluation protocols. In this way, common patterns of misalignments can be found. Furthermore, the counts in the summary show valuable data about the alignment process as for instance the alignment step in which the most incorrect or partially correct units were linked and so on. This information is very useful for future improvements and adjustments of the alignment system.

5. Sample Evaluation

5.1. Creating gold standards

The basic setup of the evaluations was to create gold standards for a majority of sub-corpora of the PLUG Corpus[2]. The PLUG Corpus consists of parallel texts of different language pairs and genres. In this paper we will focus on 3 Swedish/English sub-corpora with a size of 132,000 up to 385,000 words.

The gold standards were created by randomly generating 500 tokens occurring in different sentences from the source half of each sub-corpus.

[2] An overview of the PLUG corpus can be found at the PLUG home page (http://stp.ling.uu.se/~corpora/plug/).

5.2. Applying the Word Alignment System

The alignment system was applied to all bitexts in the PLUG corpus. The parameter settings of the system were adjusted by multiple test alignments according to the values that could be measured by the automatic evaluation using the gold standards (Ahrenberg et al. 1999). Table 1 gives an overview of three sample alignments of Swedish/English bitexts from the PLUG corpus. All the three experiments applied prior phrase generation for both languages.

genre	size (ca)	$\text{precision}_{\text{PWA}}$	$\text{recall}_{\text{PWA}}$	F
technical	385,000	81.26	64.47	71.90
fiction	132,000	83.53	51.61	63.80
political	180,000	69.04	41.44	51.76

Table 3. Word alignment results of three Swedish/English bitexts.

The results in table 3 show clearly the relation between the performance of the alignment software and the type of text under consideration. The largest outcome could be achieved for the technical sub-corpus. Here, the alignment system profited mainly from the large amount of short statements and direct translations that use a strict technical terminology. The performance on the fictional text is quite similar in precision compared to the experiment with the technical text. However, the recall value dropped significantly. The worst result was yielded for the sub-corpus of political texts. This can be explained by different facts. First, freely translated sections can be found frequently. The number of 'null-links' (about 10%) in the gold standard for this text is one measurable reflection of this fact. Another reason for the poor performance might be related to the unspecified translation history of this corpus. It is neither known which part of the bitext should be considered to be the origin nor if there was another intermediate language involved in the translation process.

5.3. Evaluation and the extracted lexicon

When testing different setups of PWA, it was found that configurations where all the modules and tests were used, increased the number of type links (i.e., the size of the extracted lexicon) by more than 300% compared to when only the statistical core was used. In table 4 below this is illustrated by comparing the size of extracted lexicons made by the baseline configuration (BASE) and by the ALL configuration one of the sub-corpora.

	Size of extracted lexicon	
genre	BASE	ALL
fiction	2,445	8,639

Table 4. The size of extracted bilingual lexicons.

The fact that the number of link types increases drastically when all the modules are invoked does not stand out clearly when configurations are compared to a randomly generated gold standard. For example, the automatically calculated $\text{recall}_{\text{PLUG}}$ score for the fictional text was 63.0% (BASE) and 74.4% (ALL). The

differences are not to be found in the actual links made by the system but by the way they are measured. High type recall usually means that a system is better at linking low-frequency items, but in order to capture the characteristics of a certain system, it is necessary to vary the strategies for creating samples, or to complement evaluations using randomized gold standards with other methods.

All the gold standards used in the tests were created without restrictions on frequency or categories. The same fictional text as above was also tested against different gold standards, which had been created with different sampling methods. One type of gold standard was made with a frequency-balanced approach (100 entries with frequency 1-2, 100 with frequency 3-4, 100 with frequency 5-9, 100 with frequency 10-40 and 100 with frequency above 40). The other type of gold standard was also frequency-balanced but contained only content words as input words for the annotation. The results from comparing the system output to these different types of gold standards are illustrative:

	genre: fiction	
gold standard type	**recall$_{PLUG}$**	**precision$_{PLUG}$**
A. Random text tokens	74.4	81.5
B. Frequency-balanced	69.0	85.6
C. Frequency balanced content words	64.0	87.1

Table 5. Recall and precision for the ALL configuration as evaluated by three different gold standards

As can be expected, the selection of content words made recall decrease and precision increase. Note that in spite of the large differences the recall and precision data in table 5 are taken from a single execution of PWA for each text. This means that the sampling strategy used when reference data are created, has a (surprisingly) great impact on the figures for recall and precision. Thus, a system's results when compared to a gold standard containing links that have been collected according to some given criteria, must not be generalised beyond that class of links.

6. Conclusions

In this paper, the evaluation problem of word alignment systems has been addressed.

We recommend the usage of prior references for the evaluation of word alignment systems, in spite of efforts needed to create gold standards, annotation tools and annotation guidelines.

Within the PLUG project a system (PLA) and a set of guidelines for the creation of English-Swedish reference links and an automatic scorer for measuring system results have been developed as well as measures for recall and precision for full text alignment. Evaluation metrics should in our view consider both halves of the links. The proposed PWA measures give credit to systems that match reference links closely on both the source and target side.

Furthermore, measures of recall and precision must be complemented with information about sampling strategies and word types covered (of course in addition to information about text type and language pair).

The PLUG Word Aligner (PWA) including the PWA Scorer is available for academic research purposes. The system is provided for Linux, Sun, and Windows platforms.

7. References

Ahrenberg, L., M. Merkel, Sågvall Hein, A., and Tiedemann, J. (1999). Evaluation of LWA and UWA. PLUG deliverable 3A.1. Internal report.

Ahrenberg, L., Andersson, M. and Merkel, M., 1998. A simple hybrid aligner for generating lexical correspondences from parallel texts. In *Proceedings of COLING-ACL'98*, Montreal, Canada, 1998, pp. 29-35.

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

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

Merkel, M. and Ahrenberg, L., 1999. Evaluating word alignment systems. PLUG Report.

Merkel, M., Andersson, M., and Ahrenberg, L, forthcoming. The PLUG Link Annotator - Interactive Construction of Data from Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Merkel, M., (1999). Annotation Style Guide for the PLUG Link Annotater. Linköping. PLUG report, Linköping University.

Merkel, M. and Andersson, M., forthcoming. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. To appear in *Proceedings from RIAO*, 2000, Paris.

Sågvall Hein, A., forthcoming. The PLUG Project: Parallel corpora in Linköping, Uppsala, Göteborg: Aims and achievements. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Tiedemann, J., (1997). Automatical Lexicon Extraction from Aligned Bilingual Corpora. Diploma thesis, Otto-von-Guericke-University, Magdeburg, Department of Computer Science.

Tiedemann, J. (1998). Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen.

Tiedemann, J., forthcoming. Uplug - A Modular Corpus Tool for Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University, Sweden.

Tiedemann, J., forthcoming. Word Alignment - Step by Step In *Proceedings of the 12th Nordic Conference on Computational Linguistics NODALI99*, University of Trondheim/Norway 1999, to appear.

Véronis, J. and Langlais, P., forthcoming. Evaluation of parallel text alignment system - The ARCADE project. To be published in *Parallel Text Processing*. J. Véronis. Berlin, Kluwer.