Cross-corpus evaluation of word alignment

Sylwia Ozdowska

National Centre for Language Technology, School of Computing
Dublin City University, Glasnevin, Dublin 9, Ireland
sozdowska@computing.dcu.ie

Abstract

We present the procedures we implemented to carry out system oriented evaluation of a syntax-based word aligner —Alibi. We take the approach of regarding cross-corpus evaluation as part of system oriented evaluation assuming that corpus type may impact alignment performance. We test our system on three English–French parallel corpora. The evaluation procedures include the creation of a reference set with multiple annotations of the same data for each corpus, the assessment of inter-annotator agreement rates and an analysis of the reference sets. We show that alignment performance varies across corpora according to the multiple references produced and further motivate our choice of preserving all reference annotations without solving disagreements between annotators.

1. Introduction

Depending on the stage of its life cycle, the performance of an NLP system can be assessed through system oriented evaluation, task oriented evaluation or user oriented evaluation (Hirschman and Mani, 2003; Paroubek, 2004). This paper focuses on a system oriented evaluation experience set up to monitor the performance of a syntax-based word alignment system —Alibi (Ozdowska, 2006)—throughout its development. System oriented evaluation aims to assess a system's intrinsic potential as a technology irrespective of its capabilities as a real-world operational application (Chaudiron, 2004).

Intrinsic performance is evaluated using standard metrics such as precision, recall and f-measure, by comparing a system's output to human-annotated reference data. For word alignment, output and reference data consist of word pairs that are (supposed to be) mutual translations within pairs of aligned sentences. The reliability of reference data can be maximised through the use of an annotation guide and the creation of multiple annotations for the same data (Melamed, 1998b; Mihalcea and Pedersen, 2003; Véronis and Langlais, 2000; Chiao et al., 2006).

Word alignment systems are basically evaluated on one particular type of corpus. Cross-corpus evaluation is still relatively rare in NLP (Kilgarriff and Grefenstette, 2003) probably because it is difficult to set up. Nevertheless, evaluating NLP systems from a cross-corpus perspective is crucial as it makes it possible to assess the influence of corpus type on performance. Concerning Alibi, cross-corpus evaluation was regarded as part of system oriented evaluation. Our hypothesis was that the granularity of alignments and the level of syntactic correspondence depend on corpus type; our objective was to assess how this impacts on alignment quality.

The reminder of this paper is organised as follows. First (section 2.) we briefly present Alibi (section 2.1.) and the corpora used to evaluate it (section 2.2.). Then (section 3.) we describe the evaluation procedures we set up: creation of reference sets (section 3.1.), assessment of inter-annotator agreement rates (section 3.2.) and analysis of the reference sets (section 3.3.). In sections 4. and 5. we study and discuss the results obtained, and conclude in section 6.

2. Context

2.1. System

Alibi is a rule-based word alignment system. It has been developed according to the following analogy-based hypothesis formulated in (Debili and Zribi, 1996): if there is a pair of words that are mutual translations within aligned sentences (i.e. *anchor words* such as *Community* and *Communauté* in figure 1) then the translational equivalence link (alignment link) can be projected to syntactically connected words (*ban* and *a interdit* in figure 1).

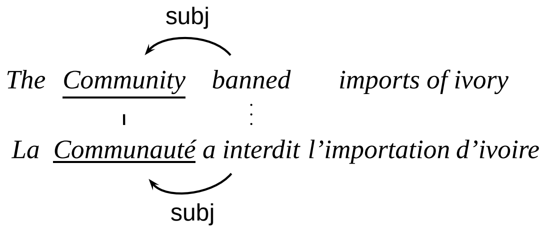
subj
The Community banned imports of ivory
La Communauté a interdit l'importation d'ivoire
subj

Figure 1: Syntax-based alignment

Anchor words are obtained based on statistical models and/or cognates. Dependency relations are identified with the Syntex parser (Bourigault et al., 2005) in both languages. Alignment links are projected through a set of syntactic alignment rules manually defined according to the following pattern (where H is the POS of the head word, rel the label indicating the nature of the dependency and DEP the POS of the dependent): H_e–rel_e–DEP_e / H_f–rel_f–DEP_f. For example, looking back to Figure 1, the rule used to project the alignment link is: V–subj–$\underline{\mathsf{N}}$ / V–subj–$\underline{\mathsf{N}}$.

2.2. Corpora

Alibi was tested on three English/French parallel corpora aligned at the sentence level and parsed with Syntex: INRA, JOC and HANSARD.

Inra is a corpus of research and popular science articles on agronomics containing about 300,000 word tokens and 7,137 aligned sentences, with an average sentence length of 19.2 words for English and 21.4 words for French. Inra was collected at the French National Research Institute for Agronomics[1].

[1] The corpus was provided by A. Lacombe from INRA.

JOC is a corpus of texts issued by the European Commission containing about 300,000 word tokens and 8,759 aligned sentences with an average sentence length of 23 words for English and 27.2 words for French. JOC was provided within the framework of the ARCADE compaigns (Véronis and Langlais, 2000; Chiao et al., 2006)[2].

HANSARD is a corpus of Canadian parliamentary debates containing about 250,000 word tokens and 8,000 aligned sentences with an average sentence length of 15 words for English and 16.6 words for French. The Hansards have been widely exploited in alignment and statistical machine translation, e.g. they were used for word alignment system evaluation in the HLT-NAACL'03 campaign (Mihalcea and Pedersen, 2003)[3].

3. Evaluation procedures

3.1. Human annotation

For the English–French language pair, there exist several reference sets for word alignment that are built out of corpora such as the Bible (Melamed, 1998a), the Hansards (Och and Ney, 2003) or JOC (Véronis and Langlais, 2000; Kraif, 2001). However, these reference sets were produced according to different manual annotation schemes and hence they do not constitute a well-suited resource for the purposes of cross-corpus evaluation.

To investigate whether corpus type affects alignment performance, we had to create a reference set for each of the copora used in our experiments. To do this, we relied on state-of-the-art word alignment evaluation principles: annotation guidelines were established and multiple annotations of the same data were produced in order to increase the reliability of the reference sets. The underlying motivation behind the definition of annotation guidelines is to guarantee that the annotation is as consistent as possible both internally and externally, i.e. that the same decisions are made by each annotator regarding different occurrences of the same type of bilingual configuration, and also that the same decisions are made by different annotators. Setting up an annotation strategy to avoid inconsistencies seems to be all the more relevant as the annotators who take part in the manual annotation task are not necessarily familiar with alignment and/or translation.

Three annotators (J1, J2 and J3) contributed to the annotation task. A sample of 120 sentences was extracted from each corpus. 60 sentences of each sample were aligned by one annotator and 60 were aligned by two annotators: 20 of them were aligned by J1 and J2, 20 by J2 and J3 and 20 by J1 and J3.

3.2. Inter-annotator agreement

Human annotation is to some extent subjective as it depends on individual interpretations that may vary from one person to another, all the more if annotation of translational correspondences within aligned sentences is considered. At subsentential level, correspondences are fuzzy and not necessarily straightforward and it is often difficult even for a human to determine which source word or sequence of words corresponds to which target word or sequence of words (Och and Ney, 2003).

Reliability of the reference sets was measured by computing inter-annotator agreement rates, which made it possible to guarantee that reference alignments were consistent enough across annotators. Given two reference sets X and Y containing human annotations of the same data, *i.e.* pairs of source words $(u, v) \in X$ and pairs of target words $(u, v) \in Y$, inter-annotator agreement rate is estimated comparing the pairs $(u, v) \in X$ and the pairs $(u, v) \in Y$ in order to determine the proportion of common pairs as compared to all pairs in Y on one hand ($A_{X/Y}$), and as compared to all pairs in X on the other hand ($A_{Y/X}$).

$$A_{X/Y} = \frac{\text{nb of common alignments}}{\text{nb of alignments} \in \text{Y}}$$

$$A_{Y/X} = \frac{\text{nb of common alignments}}{\text{nb of alignments} \in \text{X}}$$

The inter-annotator agreement rate combines both values $A_{X/Y}$ and $A_{Y/X}$, for example through a harmonic mean:

$$A = \frac{2A_{Y/X}A_{X/Y}}{A_{Y/X} + A_{X/Y}}$$

The rates were computed for each pair of annotators and each sample. The results are shown in Table 1. The overall agreement between pairs of annotators is reasonably high: over 0.7. The agreement rates are fairly stable across pairs of annotators, although they vary across corpora. Considering each pair of annotators individually, the inter-annotator agreement rate is much lower on HANSARD, 0.77 on average, than on INRA and JOC, respectively 0.89 and 0.86 on average.

	AJ1J2	AJ1J3	AJ2J3
INRA	0.90	0.89	0.88
JOC	0.87	0.86	0.85
HANSARD	0.76	0.82	0.72

Table 1: Inter-annotator agreement rates

The human annotations mainly differed due to the granularity of alignments. For example, the annotation is chunk based in sentence (1)[4], *i.e.* the chunk *allis shad* is linked to the chunk *grande alose*, whereas it is word based in sentence (2), *i.e.* *allis* is linked to *grande* on one hand and *shad* is linked to *alose* on the other hand.

(1) *The **[allis shad]**$_1$ **[is_considered to be]**$_2$ a vulnerable species*
*La **[grande alose]**$_1$ **[est considérée comme]**$_2$ une espèce vulnérable*

[2] It is distributed by ELRA/ELDA (Evaluations and Language Resources Distribution Agency, www.elda.org).

[3] The corpus used in the experiments described in this paper was provided by RALI (Laboratoire de Recherche Appliquée en Linguistique Informatique).

[4] The underscore indicates deviations from standard tokenisation resulting from the parser's pre- and post-processing procedures: *is_considered* corresponds to one token.

(2) *The* ***allis***$_1$ ***shad***$_2$ ***is_considered***$_3$ ***[to be]***$_4$ *a vulnerable species*
La ***grande***$_1$ ***alose***$_2$ ***[est considérée]***$_3$ ***comme***$_4$ *une espèce vulnérable*

3.3. Types of correspondences

A detailed analysis of the reference sets allowed us to observe the distribution of translational correspondences according to their type (Table 2): 1–1 (e.g. *disease*/*maladie*), null (source or target word(s) without correspondence, chunk (several source and/or target words involved in the correpondence, e.g. *about*/*au sujet de*).

From Table 2, we can see that the reference set for INRA is the one with the highest rate of 1–1 correspondences, meaning that the translation is mostly literal. Conversely, the hightest rate of chunk correspondences is found in HANSARD. This time, the translation appears to be mostly free which may be due to the nature of the data, i.e. speech data. Upon looking at inter-annotator agreement, we noted that the rate was lower on HANSARD. The additional information about the distribution of correspondences seems to indicate that the chances that the annotations diverge across annotators increase as the proportion of 1–1 correspondences drops; different boundaries tend to be chosen to delimit corresponding chunks.

	1–1	**null**	**chunk**
	INRA		
J1	58%	18%	24%
J2	64%	15%	21%
J3	57%	13%	30%
	JOC		
J1	55%	25%	20%
J2	51%	22%	27%
J3	53%	21%	26%
	HANSARD		
J1	39%	19%	42%
J2	43%	21%	36%
J3	45%	25%	30%

Table 2: Distribution of correspondences accroding to their type

4. Results

The annotations produced by each annotator for each corpus were merged to get three reference sets of 180 sentences each (corresponding to 120 different sentences), that is to say all the annotations were kept for evaluation purposes. The performance of the ALIBI system was evaluated against the reference sets using standard evaluation metrics: precision (P), recall (R) and f-measure (F).

$$P = \frac{\text{correct output alignments}}{\text{output alignments}}$$

$$R = \frac{\text{correct output alignments}}{\text{output alignments}}$$

$$F = \frac{2PR}{P+R}$$

The objective was to consider the value of incorporating syntax into the alignment process and to assess the impact of corpus type on alignment quality. The experimental results are given in Table 3. The alignment performance obtained with ALIBI on each of the three corpora are compared to a baseline consisting of the intersection of Giza++ IBM 4 alignments in both source-to-target and target-to-source directions (Och and Ney, 2003). The baseline alignment corresponds to the anchor alignments in the experiments reported in this paper. In addition to P, R and F, absolute and relative contributions are also shown (absolute / relative).

Globally, the results are satisfactory. ALIBI improves upon the baseline across all three corpora giving absolute increases in F of between 0.04 and 0.06 (relative increases are between 0.05 and 0.10). The absolute increases are relatively stable across the three corpora. Conversely, looking to the relative increases, we note that the gain is twice as high for HANSARD (0.10) as it is for INRA and JOC (0.05). The situation is similar when increases in P and R are considered, i.e. increases are significantly higher on HANSARD than on INRA and JOC for these measures. ALIBI achieves a broader coverage, absolute increases in R are between 0.09 and 0.10 (relative increases are between 0.13 and 0.23), but yields slight absolute decreases in P of between 0.04 and 0.07 (between 0.04 and 0.08 relative decrease). New alignments are induced based on the syntactic projection rules. However, not all of them are correct since errors arising from the automation of the whole process, in particular parsing errors and achoring errors, are unavoidable[5] and may have a further negative impact on the alignment process. On the other hand, some of the alignment errors are due to rephrasings that are made during translation.

Finally, looking to the P, R and F scores, we observe considerable differences according to the input corpus. ALIBI performs significantly better on INRA (0.91 P and 0.75 R) and JOC (0.87 P and 0.67 R) than on HANSARD (0.82 P and 0.53 R). There is a clear-cut variation in performance when comparing INRA and JOC *vs.* HANSARD.

5. Discussion

The decision to preserve all the annotations for evaluation purposes and not to solve disagreements between annotators can be motivated as follows. First, as previously stated, inter-annotator agreement rates on all three reference sets are reasonably high, meaning that most of the annotations are similar across annotators. While manual annotation guidelines aim to minimize disagreement between annotators, the annotation ultimately depends on annotators' individual assessment of each bilingual configuration they have to process. Bearing in mind that translational correspondences are fuzzy and hence may be difficult to make explicit, it seems reasonable to admit that different annotations of the same data may co-exist. In other words, there

[5] A preliminary evaluation carried out directly on the output alignments in order to evaluate individually each alignment rule showed that 60% of alignment errors were due to a syntactic analysis error when considering the rule that projects alignment links from subject achor pairs to verbs.

	INRA		JOC		HANSARD	
	Base	**ALIBI**	**Base**	**ALIBI**	**Base**	**ALIBI**
P	0.95	0.91 (-0.04 / -0.04)	0.93	0.87 (-0.06 / -0.06)	0.89	0.82 (-0.07 / -0.08)
R	0.66	0.75 ($+0.09$ / $+0.13$)	0.58	0.67 ($+0.09$ / $+0.15$)	0.43	0.53 ($+0.10$ / $+0.23$)
F	0.78	0.82 ($+0.04$ / $+0.05$)	0.71	0.75 ($+0.04$ / $+0.06$)	0.58	0.64 ($+0.06$ / $+0.10$)

Table 3: Performance of ALIBI

might be more than one plausible annotation and the existence of multiple annotations can thus be seen as the reflection of the complexity of both translation and alignment processes. An example of multiple annotation of sentence (3) is given in (3-a) and (3-b). English words or sequences of words are indexed with French words or sequences of words they are aligned to. Words that are neither indexed nor used as index are aligned to null.

(3) *I not was_asking for a detailed explanation as to what he was doing*
Je ne lui ai pas demandé de me fournir de telles explications sur ces activités

a. ***I***$_{\textit{je}}$ ***not***$_{\textit{(ne pas)}}$ ***[was_asking for]***$_{\textit{(ai demandé)}}$
[a detailed]$_{\textit{(de telles)}}$ ***explanation***$_{\textit{explications}}$
[as to what he was doing]$_{\textit{(sur ces activités)}}$

b. ***I***$_{\textit{je}}$ ***not***$_{\textit{(ne pas)}}$ ***was_asking***$_{\textit{(ai demandé)}}$
for$_{\textit{(de me fournir)}}$ ***a detailed explanation***$_{\textit{explications}}$
[as to]$_{\textit{sur}}$ ***[what he was doing]***$_{\textit{(ces activités)}}$

Regarding the results, the precise reasons why performance varies across corpora are not easy to identify. The distribution of correspondences according to their type in each reference set is probably a clue to the cause of variability. First of all, each corpus displays a different degree of 1–1 correspondences. On the other hand, correspondence at the syntactic level does not necessarily imply correspondence at the semantic level. Finally chunk correspondences of which the highest rate is found in HANSARD lead to an overgeneration of reference alignments since they are split into individual links between each source word and each target word contained in the corresponding chunks. It is not clear to what extent the precision and recall scores obtained through this method account for the actual accuracy and coverage of the achieved alignment.

6. Conclusion

Building reference sets is time consuming especially since several corpora and several annotators are involved. Reference sets for word alignment are generally quite small and one wonders whether their size is appropriate for reliable evaluation results to be obtained. There is no question of building either extensive or exhaustive reference sets but rather determining an optimal size to make sure that estimated performance is stable, meaning it would not vary significantly if more reference data was used. A set of 180 aligned sentences is probably not enough reference data in this respect.

7. References

Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques, and Sylwia Ozdowska. 2005. SYNTEX, analyseur syntaxique de corpus. In *EASy Workshop, Conférence Traitement Automatique des Langues Naturelles*, Dourdan, France.

Stéphane Chaudiron. 2004. La place de l'usager dans l'évaluation des systèmes de recherche d'informations. In Stéphane Chaudiron, editor, *Évaluation des systèmes de traitement de l'information*, pages 287–310. Hermès, Paris.

Yun-Chuang Chiao, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, and Wajdi Zaghouani. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. In *5^{th} Conference on Language Resources and Evaluation (LREC'06)*, pages 1975–1978, Genoa, Italy.

Fathi Debili and Adnane Zribi. 1996. Les dépendances syntaxiques au service de l'appariement de mots. In *$10^{ème}$ Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 81–90, Rennes, France.

Lynette Hirschman and Inderjeet Mani. 2003. Evaluation. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 414–429. Oxford University Press.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue of web as a corpus. *Computational Linguistics*, 29(3):333–338.

Olivier Kraif. 2001. *Constitution et exploitation de bitextes pour l'aide à la traduction*. Phd. Thesis, University Nice Sophia Antipolis, France.

I. Dan Melamed. 1998a. Annotation style guide for the blinker project. Technical repport, Institute for Research in Congnitive Science, University of Pennsylvania, Philadelphia, USA.

I. Dan Melamed. 1998b. Manual annotation of translational equivalence: The blinker project. Technical repport, Institute for Research in Congnitive Science, University of Pennsylvania, Philadelphia, USA.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Canada.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 1(29):19–51.

Sylwia Ozdowska. 2006. ALIBI, *un système d'ALIgnement BIlingue à base de règles de propa-*

gation syntaxique. Phd. Thesis, University Toulouse-Le Mirail, France.

Patrick Paroubek. 2004. L'évaluation des systèmes d'analyse morphosyntaxique et syntaxique. In Stéphane Chaudiron, editor, *Évaluation des systèmes de traitement de l'information*, pages 101–125. Hermès, Paris.

Jean Véronis and Philippe Langlais. 2000. Evaluation of parallel text alignment systems. The ARCADE project. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, chapter 19, pages 369–388. Kluwer Academic Publishers, Dordrecht.

Acknowledgements

Thanks to Science Foundation Ireland (`http://www.sfi.ie`) Principal Investigator Award 05/IN/1732 for part-funding this research.