Using bilingual dependencies to align words in Enlish/French parallel corpora

Sylwia Ozdowska
ERSS - CNRS & Université de Toulouse le Mirail
5 allées Antonio Machado
31058 Toulouse Cedex France
ozdowska@univ-tlse2.fr

Abstract

This paper describes a word and phrase alignment approach based on a dependency analysis of French/English parallel corpora, referred to as alignment by "syntax-based propagation." Both corpora are analysed with a deep and robust dependency parser. Starting with an anchor pair consisting of two words that are translations of one another within aligned sentences, the alignment link is propagated to syntactically connected words.

1 Introduction

It is now an acknowledged fact that alignment of parallel corpora at the word and phrase level plays a major role in bilingual linguistic resource extraction and machine translation. There are basically two kinds of systems working at these segmentation levels: the most widespread rely on statistical models, in particular the IBM ones (Brown *et al*., 1993); others combine simpler association measures with different kinds of linguistic information (Arhenberg *et al*., 2000; Barbu, 2004). Mainly dedicated to machine translation, purely statistical systems have gradually been enriched with syntactic knowledge (Wu, 2000; Yamada & Knight, 2001; Ding *et al*., 2003; Lin & Cherry, 2003). As pointed out in these studies, the introduction of linguistic knowledge leads to a significant improvement in alignment quality.

In the method described hereafter, syntactic information is the kernel of the alignment process. Indeed, syntactic dependencies identified on both sides of English/French bitexts with a parser are used to discover correspondences between words. This approach has been chosen in order to capture frequent alignments as well as sparse and/or corpus-specific ones. Moreover, as stressed in previous research, using syntactic dependencies seems to be particularly well suited to coping with the problem of linguistic variation across languages (Hwa *et al*., 2002). The implemented procedure is referred to as "syntax-based propagation".

2 Starting hypothesis

The idea is to make use of dependency relations to align words (Debili & Zribi, 1996). The reasoning is as follows (Figure 1): if there is a pair of anchor words, i.e. if two words $w1_i$ (*community* in the example) and $w2_m$ (*communauté*) are aligned at the sentence level, and if there is a dependency relation between $w1_i$ (*community*) and $w1_j$ (*ban*) on the one hand, and between $w2_m$ (*communauté*) and $w2_n$ (*interdire*) on the other hand, then the alignment link is propagated from the anchor pair (*community*, *communauté*) to the syntactically connected words (*ban*, *interdire*).

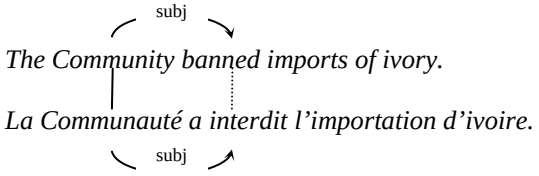

Figure 1. Syntax-based propagation

Proceedings of the ACL Student Research Workshop, pages 127–132,
Ann Arbor, Michigan, June 2005.

We describe hereafter the overall design of the syntax-based propagation process. We present the results of applying it to three parsed English/French bitexts and compare them to the baseline obtained with the giza++ package (Och & Ney, 2000).

3 Corpora and parsers

The syntax-based alignment was tested on three parallel corpora aligned at the sentence level: INRA, JOC and HLT. The first corpus was compiled at the National Institute for Agricultural Research (INRA)[1] to enrich a bilingual terminology database used by translators. It comprises 6815 aligned sentences[2] and mainly consists of research papers and popular-science texts.

The JOC corpus was made available in the framework of the ARCADE project, which focused on the evaluation of parallel text alignment systems (Veronis & Langlais, 2000). It contains written questions on a wide variety of topics addressed by members of the European Parliament to the European Commission, as well as the corresponding answers. It is made up of 8765 aligned sentences.

The HLT corpus was used in the evaluation of word alignment systems described in (Mihalcea & Pederson, 2003). It contains 447 aligned sentences from the Canadian Hansards (Och & Ney, 2000).

The corpus processing was carried out by a French/English parser, SYNTEX (Fabre & Bourigault, 2001). SYNTEX is a dependency parser whose input is a POS tagged[3] corpus — meaning each word in the corpus is assigned a lemma and grammatical tag. The parser identifies dependencies in the sentences of a given corpus, for instance subjects and direct and indirect objects of verbs. The parsing is performed independently in each language, yet the outputs are quite homogeneous since the syntactic dependencies are identified and represented in the same way in both languages.

In addition to parsed English/French bitexts, the syntax-based alignment requires pairs of anchor words be identified prior to propagation.

4 Identification of anchor pairs

To derive a set of words that are likely to be useful for initiating the propagation process, we implemented a widely used method of co-occurrence counts described notably in (Gale & Church, 1991; Ahrenberg *et al.*, 2000). For each source (*w1*) and target (*w2*) word, the Jaccard association score is computed as follows:

$$j(w1, w2) = f(w1, w2)/f(w1) + f(w2) – f(w1, w2)$$

The Jaccard is computed provided the number of overall occurrences of *w1* and *w2* is higher than 4, since statistical techniques have proved to be particularly efficient when aligning frequent units. The alignments are filtered according to the j(*w1*, *w2*) value, and retained if this value was 0.2 or higher. Moreover, two further tests based on cognate recognition and mutual correspondence condition are applied.

The identification of anchor pairs, consisting of words that are translation equivalents within aligned sentences, combines both the projection of the initial lexicon and the recognition of cognates for words that have not been taken into account in the lexicon. These pairs are used as the starting point of the propagation process[4].

Table 1 gives some characteristics of the corpora.

	INRA	JOC	HLT
aligned sentences	6815	8765	477
anchor pairs	4376	60762	996
w1/source sentence	21	25	15
w2/target sentence	24	30	16
anchor pairs/sentence	6.38	6.93	2.22

Table 1. Identification of anchor pairs

5 Syntax-based propagation

5.1 Two types of propagation

The syntax-based propagation may be performed in two different directions, as a given word is likely to be both governor and dependent with respect to other words. The first direction starts with dependent anchor words and propagates the alignment link to the governors (Dep-to-Gov propagation). The Dep-to-Gov propagation is *a priori* not ambiguous since one dependent is governed at

1 We are grateful to A. Lacombe who allowed us to use this corpus for research purposes.

2 The sentence-level alignment was performed using Japa (http://www.rali.iro.umontreal.ca).

3 The French and English versions of Treetagger (http://www.ims.uni-stuttgart.de) are used.

4 The process is not iterative up to date so the number of words it allows to align depends on the initial number of anchor words per sentence.

most by one word. Thus, there is just one relation on which the propagation can be based. The second direction goes the opposite way: starting with governor anchor words, the alignment link is propagated to their dependents (Gov-to-Dep propagation). In this case, several relations that may be used to achieve the propagation are available, as it is possible for a governor to have more than one dependent. So the propagation is potentially ambiguous. The ambiguity is particularly widespread when propagating from head nouns to their nominal and adjectival dependents. In Figure 2, there is one occurrence of the relation pcomp in English and two in French. Thus, it is not possible to determine *a priori* whether to propagate using the relations mod/pcomp2, on the one hand, and pcomp1/pcomp2', on the other hand, or mod/pcomp2' and pcomp1/pcomp2. Moreover, even if there is just one occurrence of the same relation in each language, it does not mean that the propagation is of necessity performed through the same relation, as shown in Figure 3.

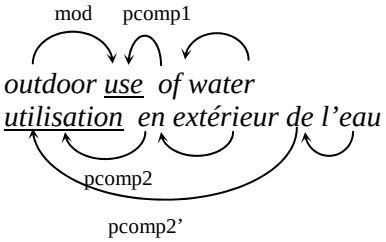

Figure 2. Ambiguous propagation from head nouns

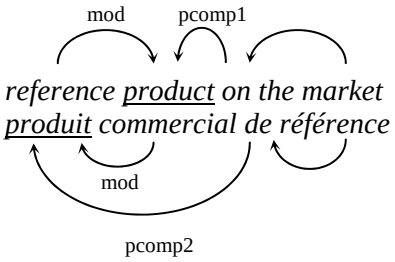

Figure 3. Ambiguous propagation from head nouns

In the following sections, we describe the two types of propagation. The propagation patterns we rely on are given in the form CDep-rel-CGov, where CDep is the POS of the dependent, rel is the dependency relation and CGov, the POS of the governor. The anchor element is underlined and the one aligned by propagation is in bold.

5.2 Alignment of verbs

Verbs are aligned according to eight propagation patterns.

DEP-TO-GOV PROPAGATION TO ALIGN GOVERNOR VERBS. The patterns are: Adv-mod-**V** (1), N-subj-**V** (2), N-obj-**V** (3), N-pcomp-**V** (4) and V-pcomp-**V** (5).
*(1) The net **is** then **hauled** to the shore.*
*Le filet **est** ensuite **halé** à terre.*
*(2) The fish **are** generally **caught** when they migrate from their feeding areas.*
*Généralement les poissons **sont capturés** quand ils migrent de leur zone d'engraissement.*
*(3) Most of the young shad **reach** the sea.*
*La plupart des alosons **gagne** la mer.*
*(4) The eggs are very small and **fall** to the bottom.*
*Les oeufs de très petite taille **tombent** sur le fond.*
*(5) X is a model which **was designed** to stimulate…*
*X est un modèle qui **a été conçu** pour stimuler…*

GOV-TO-DEP PROPAGATION TO ALIGN DEPENDENT VERBS. The alignment links are propagated from the dependents to the verbs using three propagation patterns: **V**-pcomp-V (1), **V**-pcomp-N (2) and **V**-pcomp-Adj (3).
*(1) Ploughing tends to **destroy** the soil microaggregated structure.*
*Le labour tend à **rompre** leur structure microagrégée.*
*(2) The capacity to **colonize** the digestive mucosa…*
*L'aptitude à **coloniser** le tube digestif…*
*(3) An established infection is impossible to **control**.*
*Toute infection en cours est impossible à **maîtriser**.*

5.3 Alignment of adjectives and nouns

The two types of propagation described in section 5.2 for use with verbs are also used to align adjectives and nouns. However, these latter categories cannot be treated in a fully independent way when propagating from head noun anchor words in order to align the dependents. The syntactic structure of noun phrases may be different in English and French, since they rely on a different type of composition to produce compounds and on the same one to produce free noun phrases. Thus, the potential ambiguity arising from the Gov-to-Dep propagation from head nouns mentioned in section 5.1

may be accompanied by variation phenomena affecting the category of the dependents. For instance, a noun may be rendered by an adjective, or vice versa: *tax treatment profits* is translated by *traitement fiscal des bénéfices*, so the noun *tax* is in correspondence with the adjective *fiscal*. The syntactic relations used to propagate the alignment links are thus different.
In order to cope with the variation problem, the propagation is performed regardless of whether the syntactic relations are identical in both languages, and regardless of whether the POS of the words to be aligned are the same. To sum up, adjectives and nouns are aligned separately of each other by means of Dep-to-Gov propagation or Gov-to-Dep propagation provided that the governor is not a noun. They are not treated separately when aligning by means of Gov-to-Dep propagation from head noun anchor pairs.

DEP-TO-GOV PROPAGATION TO ALIGN ADJECTIVES. The propagation patterns involved are: Adv-mod-**Adj** (1), N-pcomp-**Adj** (2) and V-pcomp-**Adj** (3).
(1) The white cedar exhibits a very ***common*** *physical defect.*
Le Poirier-pays présente un défaut de forme très ***fréquent****.*
(2) The area presently ***devoted*** *to agriculture represents…*
La surface actuellement ***consacrée*** *à l'agriculture représenterait…*
(3) Only four plots were ***liable*** *to receive this input.*
Seulement quatre parcelles sont ***susceptibles*** *de recevoir ces apports.*

DEP-TO-GOV PROPAGATION TO ALIGN NOUNS. Nouns are aligned according to the following propagation patterns: Adj-mod-**N** (1), N-mod-**N**/N-pcomp-**N** (2), N-pcomp-**N** (3) and V-pcomp-**N** (4).
(1) Allis shad remain on the continental ***shelf****.*
La grande alose reste sur le ***plateau*** *continental.*
(2) Nature of micropollutant ***carriers****.*
La nature des ***transporteurs*** *des micropolluants.*
(3) The ***bodies*** *of shad are generally fusiform.*
Le ***corps*** *des aloses est généralement fusiforme.*
(4) ***Ability*** *to react to light.*
Capacité *à réagir à la lumière.*

UNAMBIGUOUS GOV-TO-DEP PROPAGATION TO ALIGN NOUNS. The propagation is not ambiguous when dependent nouns are not governed by a noun. This is the case when considering the following three propagation patterns: **N**-subj|obj-V (1), **N**-pcomp-V (2) and **N**-pcomp-Adj (3).
(1) The ***caterpillars*** *can inoculate the* ***fungus****.*
Les ***chenilles*** *peuvent inoculer le* ***champignon****.*
(2) The roots are placed in ***tanks****.*
Les racines sont placées en ***bacs****.*
(3) ...a fungus responsible for ***rot****.*
... un champignon responsable de la ***pourriture****.*

POTENTIALLY AMBIGUOUS GOV-TO-DEP PROPAGATION TO ALIGN NOUNS AND ADJECTIVES. Considering the potential ambiguity described in section 5.1, the algorithm which supports Gov-to-Dep propagation from head noun anchor words (*n1, n2*) takes into account three situations which are likely to occur.

First, each of *n1* and *n2* has only one dependent, respectively *dep1* and *dep2,* involving one of the mod or pcomp relation; *dep1* and *dep2* are aligned.
the drained whey
le lactosérum d'égouttage
⇒ (***drained, égouttage***)

Second, *n1* has one dependent *dep1* and *n2* several {$dep2_1$, $dep2_2$, …, $dep2_n$}, or vice versa. For each $dep2_i$, check if one of the possible alignments has already been performed, either by propagation or anchor word spotting. If such an alignment exists, remove the others (*dep1*, $dep2_k$) such that $k \neq i$, or vice versa. Otherwise, retain all the alignments (*dep1*, $dep2_i$), or vice versa, without resolving the ambiguity.
stimulant substances which are absent from…
substances solubles stimulantes absentes de…
(*stimulant*, {*soluble, stimulant, absent*})
already_aligned(*stimulant, stimulant*) = 1
⇒ (***stimulant***, ***stimulant***)

Third, both *n1* and *n2* have several dependents, {$dep1_1$, $dep1_2$, …, $dep1_m$} and {$dep2_1$, $dep2_2$, …, $dep2_n$} respectively. For each $dep1_i$ and each $dep2_j$, check if one/several alignments have already been performed. If such alignments exist, remove all the alignments ($dep1_k$, $dep2_l$) such that $k \neq i$ or $l \neq j$. Otherwise, retain all the alignments ($dep1_i$, $dep2_j$) without resolving the ambiguity.
unfair trading practices
pratiques commerciales déloyales
(*unfair*, {*commercial, déloyal*})
(*trading*, {*commercial, déloyal*})
already_aligned(*unfair, déloyal*) = 1

⇒ (***unfair***, ***déloyal***)
⇒ (***trading***, ***commercial***)

a big rectangular net, which is lowered…
un vaste filet rectangulaire immergé…
(*big*, {*vaste, rectangulaire, immergé*})
(*rectangular*, {*vaste, rectangulaire, immergé*})
already_aligned(*rectangular*, *rectangulaire*) = 1
⇒ (***rectangular, rectangulaire***)
⇒ (*big*, {*vaste, immergé*})

The implemented propagation algorithm has two major advantages: it permits the resolution of some alignment ambiguities, taking advantage of alignments that have been previously performed. This algorithm also allows the system to cope with the problem of non-correspondence between English and French syntactic structures and makes it possible to align words using various syntactic relations in both languages, even though the category of the words under consideration is different.

5.4 Comparative evaluation

The results achieved using the syntax-based alignment (sba) are compared to those obtained with the baseline provided by the IBM models implemented in the giza++ package (Och & Ney, 2000) (Table 2 and Table 3). More precisely, we used the intersection of IBM-4 Viterbi alignments for both translation directions. Table 2 shows the precision assessed against a reference set of 1000 alignments manually annotated in the INRA and the JOC corpus respectively. It can be observed that the syntax-based alignment offers good accuracy, similar to that of the baseline.

	INRA		**JOC**	
	sba	**giza++**	**sba**	**giza++**
Precision	0.93	0.96	0.95	0.94

Table 2. sba ~ giza++: INRA & JOC

More complete results (precision, recall and f-measure) are presented in Table 3. They have been obtained using reference data from an evaluation of word alignment systems (Mihalcea & Pederson, 2003). It should be noted that the figures concerning the syntax-based alignment were assessed in respect to the annotations that do not involve empty words, since up to now we focused only on content words. Whereas the baseline precision[5] for the HLT corpus is comparable to the one reported in Table 2, the syntax-based alignment score decreases. Moreover, the difference between the two approaches is considerable with regard to the recall. This may be due to the fact that our syntax-based alignment approach basically relies on isomorphic syntactic structures, i.e. in which the two following conditions are met: i) the relation under consideration is identical in both languages and ii) the words involved in the syntactic propagation have the same POS. Most of the cases of non-isomorphism, apart from the ones presented section 5.1, are not taken into account.

	HLT	
	sba	**giza++**
Precision	0.83	0.95
Recall	0.58	0.85
F-measure	0.68	0.89

Table 3. sba ~ giza++: HLT

6 Discussion

The results achieved by the syntax-based propagation method are quite encouraging. They show a high global precision rate — 93% for the INRA corpus and 95% for the JOC — comparable to that reported for the giza++ baseline system. The figures vary more from the HLT reference set. One possible explanation is the fact that the gold standard has been established according to specific annotation criteria. Indeed, the HLT project concerned above all statistical alignment systems aiming at language modelling for machine translation. In approaches such as Lin and Cherry's (2003), linguistic knowledge is considered secondary to statistical information even if it improves the alignment quality. The syntax-based alignment approach was designed to capture both frequent alignments and those involving sparse or corpus-specific words as well as to cope with the problem of non-correspondance across languages. That is why we chose the linguistic knowledge as the main information source.

[5] Precision, recall and f-measure reported by Och and Ney (2003) for the intersection of IBM-4 Viterbi alignments from both translation directions.

7 Conclusion

We have presented an efficient method for aligning words in English/French parallel corpora. It makes the most of dependency relations to produce highly accurate alignments when the same propagation pattern is used in both languages, i.e. when the syntactic structures are identical, as well as in cases of noun/adjective transpositions, even if the category of the words to be aligned varies (Ozdowska, 2004). We are currently pursuing the study of non-correspondence between syntactic structures in English and French. The aim is to determine whether there are some regularities in the rendering of specific English structures into given French ones. If variation across languages is subject to such regularities, as assumed in (Dorr, 1994; Fox, 2002; Ozdowska & Bourigault, 2004), the syntax-based propagation could then be extended to cases of non-correspondence in order to improve recall.

References

Ahrenberg L., Andersson M. & Merkel M. 2000. A knowledge-lite approach to word alignment. In Véronis J. (Ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 97-138.

Barbu A. M. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Actes de la Conférence JADT*.

Brown P., Della Pietra S. & Mercer R. 1993. The mathematics of statistical machine translation: parameter estimation. In *Computational Linguistics*, 19(2), pp. 263-311.

Debili F. & Zribi A. 1996. Les dépendances syntaxiques au service de l'appariement des mots. In *Actes du 10ème Congrès RFIA*.

Ding Y., Gildea D. & Palmer M. 2003. An Algorithm for Word-Level Alignment of Parallel Dependency Trees. In *Proceedings of the 9th MT Summit of International Association of Machine Translation.*

Dorr B. 1994. Machine translation divergences: a formal description and proposed solution. In *Computational Linguistics*, 20(4), pp. 597-633.

Fabre C. & Bourigault D. 2001. Linguistic clues for corpus-based acquisition of lexical dependencies. In *Proceedings of the Corpus Linguistic Conference*.

Fox H. J. 2002. Phrasal Cohesion and Statistical Machine Translation. In *Proceedings of EMNLP-02*, pp. 304-311.

Gale W. A. & Church K. W. 1991. Identifying Word Correspondences in Parallel Text. In *Proceedings of the DARPA Workshop on Speech and Natural Language*.

Hwa R., Resnik P., Weinberg A. & Kolak O. 2002. Evaluating Translational Correspondence Using Annotation Projection. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics*.

Lin D. & Cherry C. 2003. ProAlign: Shared Task System Description. In *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond.*

Mihalcea R. & Pedersen T. 2003. An Evaluation Exercise for Word Alignment. In *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.

Och F. Z. & Ney H., 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, 29(1), pp. 19-51.

Ozdowska S. 2004. Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora. In *COLING 04 Workshop on Multilingual Linguistic Resources*.

Ozdowska S. & Bourigault D. 2004. Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus. In *Actes des 14ème Congrès RFIA*.

Véronis J. & Langlais P. 2000. Evaluation of parallel text alignment systems. The ARCADE project. In Véronis J. (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 371-388

Wu D. 2000. Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars. In Véronis, J. (Ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 139-167.

Yamada K. & Knight K. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics*.