Tree Alignment through Semantic Role Annotation Projection

Tom Vanallemeersch

K.U.Leuven, Belgium
Centrum voor Computerlinguïstiek
E-mail: tallem@ccl.kuleuven.be

Abstract

Translation divergences are a challenge for MT and alignment. In this paper, we investigate whether an alignment method based on semantic knowledge improves over approaches for linguistically uninformed word alignment and purely syntax-based tree alignment. We annotate sentences with rolesets from PropBank and NomBank (verbal and nominal predicates and their semantic roles), and link predicates to their auxiliary words (auxiliary, modal and support verbs) using parse trees. We study two language pairs, English-French and English-Dutch. As no extensive semantic resource is available for French and Dutch, the annotation strategy we choose is cross-lingual semantic annotation projection, combined with automatic SRL. A manual evaluation of our system on an English-Dutch sample shows our system is successful at adding links for predicates to the output of a word alignment system (GIZA++) and two tree alignment systems (Lingua-Align and Sub-Tree Aligner). The performance for role linking is significantly lower, due to errors in the English or target parses.

1 Background

Translations tend to diverge from source texts, in different ways and by different causes. Some divergences (also called “translation shifts”) are caused by linguistic constraints, others by extralinguistic factors. As Habash et al. [6, p. 85] state, “a translation divergence occurs when the underlying concept or ’gist’ of a sentence is distributed over different words for different languages”. They mention several divergence types, such as categorial (change of part of speech), conflational (translation of two words by one, e.g. *dar puñaladas* ‘give stabs’ into *stab*), structural (e.g. addition of preposition to argument of verb) and thematic (switch of subject and object during translation). Tense and aspect are also expressed in divergent ways in languages, involving affixes (*mangeait*), auxiliary verbs (*has eaten*) and periphrastic constructions (*is going to eat*).

Divergences are a challenge for MT and alignment. In the case of MT, the system needs to make the right choices when generating the translation. In the case of alignment, both basic word alignment (linguistically uninformed SMT) and advanced forms of subsentential alignment like parse tree alignment have difficulties aligning divergent structures. Consider the alignment of the following English-Dutch sentence pair in the Europarl corpus (Koehn [9]) in Figure 1. For the sake of clarity, the start and end of the sentences and the child nodes of two non-terminal nodes are not shown. The word-for-word translation of the Dutch sentence is ‘a similar objection can be in-brought against’.

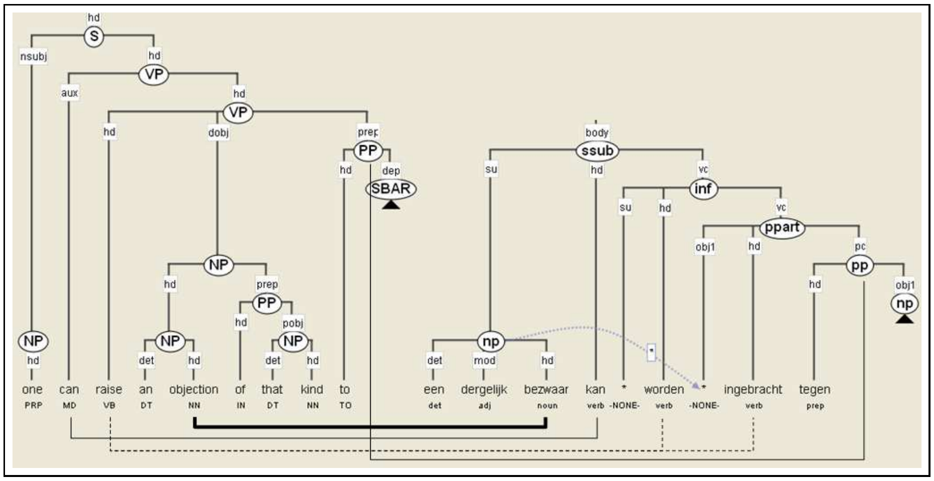

Figure 1: Alignment of divergent structures

The picture shows links created by two variants of the GIZA++ word alignment system (Och and Ney [13]), the highly precise “intersective” and the more extensive “grow-diag-final-and” variant, and by two tree alignment systems, Lingua-Align (Tiedemann and Kotzé [16]), and Sub-Tree Aligner (Zhechev and Way [18]). Only one link is established by all systems (marked in bold at the bottom of the picture). The thin solid links at the bottom of the picture are links that are only procuced by some systems, and no system produces the dashed links (one system aligns one of the words incorrectly). The dashed links involve a Dutch auxiliary of the passive (*worden*) and support verbs of the nominal predicates *objection* and *bezwaar*, which have an argument *to what is merely a report* and *tegen enkel een verslag*. The highly different morphology of the parse trees complicates tree alignment.

In order to tackle translation divergences, semantically oriented approaches have been followed in rule-based MT systems like Eurotra (Allegranza et al. [1]), for coding the argument structure of verbs and for coding tense and aspect in a language-independent way, in order to reduce the transfer step between the two languages to a minimum. In the last decade, automated semantic role labeling (SRL) using one of the available semantic frameworks has become increasingly important. The idea of semantic roles was pioneered by Fillmore [5], who posited the existence of case relations occurring at deep structure, as opposed to the surface structure of the sentence. Recently, SRL has been applied for multilingual purposes, i.e. to the domain of SMT (Wu [17]) and to parallel text annotation (Padó [14]). In this paper, we propose an approach applying semantic knowledge to the alignment of parse trees.

2 Research Question

Our research question is whether semantic knowledge can improve the alignment of words and constituents. Our assumption is that semantic knowledge is helpful in overcoming syntactic differences between sentences, thus improving over word alignment systems which use linguistically uninformed methods and over methods aligning constituents based on syntactic knowledge only.

The type of semantic knowledge we focus on consists of verbal and nominal predicates and their semantic roles, and the link between predicates and their auxiliary words. The latter are words expressing tense, aspect, modality and passive voice in case of verbal predicates, and support verbs in case of nominal predicates. We only study nominal predicates which are derived from a verb (deverbal nouns). For our purposes, we consider any verb connecting a nominal predicate to one of its arguments as a support verb (see the example *raise* - *ingebracht* in section 1). The alignment procedure we propose links predicates (and their auxiliary words) and semantic roles between sentences.

The languages we study are English, French and Dutch. This choice was motivated by the fact that English is a resource-rich language and that we want to investigate more than two languages as semantic knowledge is supposed to be applicable beyond a single language pair.

3 Method

In the following subsections, we describe the type of semantic roles we use, our procedure for annotating the predicates and semantic roles and our procedure for determining auxiliary words of a predicate.

3.1 Choice of semantic framework

There are several frameworks for semantic roles, such as FrameNet (Baker et al. [2]), VerbNet (Kipper Schuler [10]), PropBank, (Palmer et al. [15]), and NomBank (Meyers et al. [11]). They are different on many levels, such as coverage, scope of semantic roles, syntactic categories covered, and motivation for their creation. For instance, the motivation for creating PropBank was to annotate predicates and semantic roles in a full corpus, and to train a SRL system on the annotated sentences. Instead of adopting the “traditional” semantic roles (also called theta roles), such as Agent, Theme and Experiencer (which are used for instance in VerbNet), PropBank marks the roles of verbs as proto-agent (A0) or proto-patient (A1), or as A2, A3 or A4. There are also Argument Modifier roles which apply to any verb and are similar to adjuncts (e.g. AM-TMP for temporal adjunct).

There are links between the different semantic frameworks. The PropBank strategy has been applied to nouns in the NomBank project. When a noun is linked to a verb (deverbal noun), the appropriate roleset of the verb (predicate + roles) is indicated in the NomBank database. In the SemLink project (Loper et al. [7]), PropBank predicates have been linked to VerbNet theta roles, which resulted in a partial mapping between both frameworks. Note that, for alignment purposes, we consider the link of constituents with the same theta role to be stronger than if they only have the same PropBank role.

All semantic resources mentioned above have been initially created for English. Some of them are also available for other languages. For French and Dutch, no extensive resources are available; for Dutch, there exists a limited set of manually annotated PropBank annotations (Monachesi et al. [12]), a rule-based SRL system, and a SRL system trained on manually annotated data. We decided to adopt the PropBank framework for French and Dutch because of the availability of the limited Dutch resource, the framework's aptness for SRL system training, its coverage, and the fact that it covers both verbs and nouns (through NomBank). As creating an extensive PropBank resource for French and Dutch is very time-intensive, we opt for the method of cross-lingual semantic role projection, in combination with the use of a SRL system. This is the topic of the following subsections.

3.2 Projection of predicates and semantic roles

Projection of information from one language to another through alignment links was originally applied to syntactic information (from resource-rich to resource-poor language). Later on, researchers started applying it to fields such as SRL. Padó ([14]) describes an approach for English sentences manually annotated with FrameNet FEEs (frame-evoking elements, which are predicates) and roles, which projects the semantic information to German and French sentences through word alignment links from GIZA++. His primary aim is to study the degree of frame-instance parallellism across languages, i.e. to find out whether the frames used in the source sentences are preserved in the French and German sentences. A number of filters is applied in order to achieve high-precision results and diminish the influence of alignment errors.

Our approach is similar to that of Padó. We apply a SRL system to English sentences, and automated linguistic analysis to the source and target sentences, i.e. parsers that combines constituency and dependency information. We project the predicates and roles to the Dutch and French translation equivalents, using links between words produced by an alignment tool. We then filter out some projections. Our approach differs from that of Padó in the fact that we find links between predicates and their auxiliary words within one sentence and link English predicates for which no word alignment exists, based on the target parse tree and on auxiliary words of the predicate. The projection procedure is described below. The filters, the detection of auxiliary words and the linking of unaligned English predicates are described subsequently.

The projection procedure is carried out as follows:

- A source predicate is projected to a target token if all of the following conditions are fulfilled: (1) the English predicate is a verb or its roleset has a link to a verb in NomBank, (2) the predicate has at least one non-adjunct role (we ignore Argument Modifier roles for projection), (3) the tokens are linked in the word alignment, and (4) the target token is a non-auxiliary verb or a noun.

- If the source predicate was projected, each of its semantic roles is projected to the smallest constituent from the target parse tree that contains all target tokens linked to tokens from the source constituent. For instance, if the source role A1 is *this particular building* and *this* and *building* are aligned to *dit* and *gebouw*, the role A1 is projected to the constituent *dit gebouw*. We assign a weight to the projection, which is lower than 1 if some of the tokens in the target constituent don't have a link to a token in the source constituent (in the example, the weight is 1). If the weight is too low according to a given threshold, projection of the source role is cancelled.

3.3 Filters on projected information

The first filter removes a predicate (i.e. roleset) if none of its semantic roles was projected.

The second filter checks whether the verb or noun, previously annotated as a predicate through projection, has a direct syntactic connection with the constituents annotated as roles through projection. This filter targets erroneous alignments and strong translation divergences. The filter establishes the shortest path in the target parse tree between the node of the predicate and the node of the role. If none of the nodes in this path is headed by an open-class word (verb, noun, adjective or adverb)[1], the syntactic connection is considered direct. As an exception, we accept one node headed by a verb if the predicate is nominal. An example of the latter can be found in Figure 1: the path between *bezwaar* and *tegen* passes along a modal verb *kan*, an auxiliary verb *worden* and the non-auxiliary verb *ingebracht*. Note that Padó also uses a syntax-based criterion for selecting possible equivalents for a source role, i.e. “argument filtering” (p. 111).

3.4 Linking predicates to auxiliary words

In the three languages under study, there is a limited set of words expressing tense, aspect, modality and passive voice of a predicate. These words, as well as support

[1] With the exception of auxiliary and modal verbs.

verbs of a nominal predicate, are retrieved from the source tree and the target parse tree by checking the sister nodes of the verb and the direct ancestors of the verb. Verbs of modality are only retrieved if they have an infinitival complement. If both the source and target predicate have auxiliary words, we link the auxiliary words to one another. If not, the predicate in one language is aligned both to the predicate in the other language and to the latter's auxiliary word.

3.5 Linking unaligned predicates

In order to overcome weak coverage of the word alignment used for projection, we use two heuristics to link unaligned predicates:

- If a predicate has auxiliary words, and one of those words is linked to a non-auxiliary verb in the target language, we link the predicate to that verb.
- If there is a direct syntactic connection between the projection of an English role and a non-auxiliary verb in the target parse tree, this verb is linked to the predicate of the English role (unless it is already a target predicate in another roleset).

4 Resources

In this section, we describe the resources we apply as input to the method described in the previous section.

We use the Europarl corpus, as it contains translation equivalents in the three languages under study and has been completely parsed and tree-aligned for English, Dutch and French in the framework of the PaCo-MT project (`http://www.ccl.kuleuven.be/Projects/PACO/paco.php`) using Lingua-Align.

As word alignments, we use GIZA++ intersective word alignments.

The SRL system we use is the best-performing system that participated in the CoNLL 2008 task on joint learning of syntactic and semantic dependencies (Johansson and Nugues [8]). It is based on the Penn Treebank and produces syntactic output annotated with PropBank rolesets.

We use parsers which combine dependency and constituency structure:

- English: we convert the syntactic information in the output of the SRL system to Alpino XML format.
- Dutch: we use Alpino (Bouma et al. [3])
- French: we convert the output of the system described by Candito et al. ([4]) to Alpino XML format. This system is trained on a French treebank.

5 Results

We performed an evaluation of our system by running it on a sample of 100 English-Dutch sentence pairs from Europarl, and manually assessing the results. We also compared the results to the output of other alignment systems: the GIZA++ intersective and grow-diag-final-and variants (we used the word alignment produced for the whole Europarl corpus), Lingua-Align (we used the alignment produced for the PaCo-MT project) and Sub-Tree Aligner (which we ran with its standard settings). It should be noted that the Lingua-Align output is based on another English parser than the one we use for semantic projection, i.e. on the Stanford parser. We set the weight for role projection to 0.5.

The SRL system produced 347 rolesets for our sample. After projection and filtering of these rolesets by our system, 150 rolesets remained, corresponding to a total of 444 alignment links (between words or constituents). Table 1 shows the precision for each type of link: links between two predicates, between roles, between two auxiliary words and between a predicate and an auxiliary word.

	number of links	precision
predicates	146	0.95
roles	196	0.83
auxiliary words	35	0.89
predicate-auxiliary	67	0.75
total	444	0.86

Table 1: Alignment precision according to link type

In order to compare our system to the word and tree alignment systems mentioned above, we checked how many links were not present in the other systems and how many links had a different source or target part than in the other systems. Table 2 shows the number of new and different links, and (between brackets) the precision of these links. No figures are given for the role links of GIZA++, as the latter is focused on word alignment.

The precision scores of our system, as well as the comparison with other systems, indicate that our system is highly accurate when aligning predicates, creating links not existing in the other systems, or correcting links of those systems. The system performs significantly less well for roles, especially when we look at the links which are new with respect to the other systems; this is mainly due to errors of the English or target language parser (no efforts were undertaken yet to optimize the weight for role projection). The precision scores for links between predicates in one language (without auxiliary word) and an auxiliary word in the other are also significantly lower than that for predicates. However, these links are helpful in finding predicate links not present in the word alignment (see subsection 3.5).

As far as English-French is concerned, an initial evaluation of our system for that language pair points towards the same conclusions as for English-Dutch.

As an illustration of predicate linking, our system produces the following alignments:

- *you did not call me either - u heeft mij het woord niet verleend* ('you have me the word not provided'): *call* is linked both two *woord* and *verleend* (in all other systems, *call* is only linked to *woord*)
- *the competent services have not included them in the agenda (...) - de (...) diensten hebben die vragen niet op de agenda geplaatst (...)* ('the (...) services have those questions not on the agenda placed'): *included* is linked to *geplaatst* (in all other systems, it remains unlinked)

6 Conclusions and future research

In this paper, we have proposed a method for improving the alignment of words and of syntactic constituents using semantic knowledge. We aim at overcoming syntactic differences between translation-equivalent sentences by determining verbal and nominal predicates and their roles, linking auxiliary words (auxiliary, modal and support verbs) to verbal predicates, and aligning predicates, roles and auxiliary words. The semantic framework we opt for is PropBank, and the annotation strategy is cross-lingual semantic annotation projection, combined with automatic SRL.

The results of our system on a sample of English-Dutch sentences indicate that our system, which is not aiming at a full word or constituent alignment of a sentence pair, is able to improve the output of systems aiming at complete alignment, i.e. a linguistically uninformed word alignment system (GIZA++) and two tree alignment systems based on purely syntactic knowledge (Lingua-Align and Sub-Tree Aligner). Based on the links between predicates, their roles and auxiliary words, and on the information in the source and target parse tree, our system produces highly accurate links between predicates, some of which are not or incorrectly linked in the other systems. On the level of role alignment, the system is

	Lingua	Sub-Tree	GIZA++ int.	GIZA++ gdfa
new pred. links	40 (0.9)	32 (0.91)	25 (0.92)	10 (0.9)
different pred. links	6 (0.83)	12 (1)	2 (0.5)	26 (0.85)
new role links	61 (0.59)	44 (0.52)		
different role links	8 (0.5)	25 (0.8)		
new aux. words	12 (0.67)	5 (0.4)	8 (0.75)	5 (0.4)
different aux. words	1 (1)	2 (1)	1 (0)	2 (1)
new pred.-aux.	52 (0.71)	48 (0.71)	54 (0.74)	28 (0.61)
different pred.-aux.	5 (0.6)	7 (0.71)	3 (0.33)	13 (0.54)

Table 2: Comparison with other alignment approaches (number of links, precision)

less performant, being hampered by errors in the English or target language parser. While the precision scores for links between predicates in one language (without auxiliary word) and an auxiliary word in the other are also significantly lower than the scores for pairs of predicates, these links are also helpful in aligning predicates that are not included in the word alignment.

Our future research involves more extensively evaluating the system output for both language pairs through existing word alignment gold standards, optimizing the role projection threshold and training an SRL system on the annotated target sentences. By running the labeler on the same target sentences, we aim at adding new target predicates and roles to the original ones produced by the cross-lingual annotation projection. New target predicates in a sentence are aligned to source predicates based on the labels of their roles. For the evaluation, we will make use of the existing set of manually annotated PropBank rolesets for Dutch ([12]).

References

[1] Allegranza, Valerio, Bennett, Paul, Durand, Jacques, Van Eynde, Frank, Humphreys, Lee, Schmidt, Paul and Steiner, Erich H. (1991) Linguistics for Machine Translation: The Eurotra Linguistic Specifications. In Copeland, Charles, Durand, Jacques, Krauwer, Steven and Maegaard, Bente (eds.) *The Eurotra Linguistic Specifications*, Office for Official Publications of the Commission of the European Community, Luxembourg, pp. 15–123.

[2] Baker, Collin F., Fillmore, Charles J. and Lowe, John B. (1998) The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL*, Montreal, pp. 86–90.

[3] Bouma, Gosse, van Noord, Gertjan and Malouf Robert (2000) Alpino: Wide Coverage Computational Analysis of Dutch. In *Proceedings of CLIN 2000*, pp. 45–59.

[4] Candito Marie, Crabbé Benoît and Denis, Pascal (2010) Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC-2010*, La Valletta, Malta, pp. 1840–1847.

[5] Fillmore, Charles J. (1968) The Case for Case. In Bach, Emmon W. and Harms, Robert T. (eds.) *Universals in Linguistic Theory*, New York: Holt, Rinehart and Winston, pp. 1–88.

[6] Habash, Nizar and Dorr, Bonnie (2002) Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Proceedings of AMTA-2002*, Tiburon, CA, pp. 84–93.

[7] Loper, Edward, Yi, Szu-ting and Palmer, Martha (2007) Combining Lexical Resources: Mapping Between PropBank and VerbNet. In *Proceedings of IWCS*, Tilburg, the Netherlands, pp. 118–128.

[8] Johansson, Richard and Nugues, Pierre (2008) Dependency-based Semantic Role Labeling of PropBank. In *Proceedings of EMNLP*, Honolulu, Hawaii, pp. 69–78.

[9] Koehn, Philipp (2005) Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, pp. 79–86.

[10] Kipper Schuler, Karin (2005) VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania.

[11] Meyers, Adam, Reeves, Ruth, Macleod, Catherine, Szekely, Rachel, Zielinska, Veronika, Young, Brian and Grishman, Ralph (2004) Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC-04*, Lisbon, Portugal, pp. 803–806.

[12] Monachesi, Paola, Stevens, Gerwert, and Trapman, Jantine (2007) Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop, ACL Workshops*, pp. 77–84.

[13] Och, Franz Josef and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, volume 29, number 1, pp. 19–51.

[14] Padó, Sebastian (2007) Cross-Lingual Annotation Projection Models for Role-Semantic Information. Ph.D. thesis, Saarland University.

[15] Palmer Martha, Gildea, Daniel, and Kingsbury, Paul (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. In *Computational Linguistics*, 31:1, pp. 71–105.

[16] Tiedemann, Joerg and Kotzé, Gideon (2009) A Discriminative Approach to Tree Alignment. In *Proceedings of the International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning (in connection with RANLP'09 workshop)*, pp. 33–39.

[17] Wu, Dekai and Fung, Pascale (2009) Can Semantic Role Labeling Improve SMT ? In *Proceedings of EAMT 2009*, Barcelona, pp. 218–225.

[18] Zhechev, Ventsislav and Way, Andy (2008) Automatic generation of parallel treebanks. In *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing)*, pp. 1105–1112.