Automatic Constituent and Function Alignment for Parallel Treebanking

23/09, 2010

Abstract

This paper describes the development of an automatic phrase alignment method using parallel sentences parsed in Lexical-Functional Grammar as input, where similarity in analyses is used as evidence that constituents or functional elements may be linked. A set of principles for phrase alignment are formulated, based on the goals of the XPar-project [1], and an implementation is given.

1 Introduction

Lexical-Functional Grammar (LFG) is a grammatical framework where a sentence is analysed as having both a constituent structure (c-structure) and functional structure (f-structure). The former is similar to traditional phrase structure trees, while the latter is an attribute-value matrix/graph which represents dependency relations between syntactic functions (subject, object, etc.), in addition to the grammatical features of these. The argument structure of predicates is embedded in the f-structure representation.

As part of the XPar-project, we are developing a parallel treebank which will include links between corresponding constituents, as well as between corresponding syntactic functions. By utilising the information available in each monolingual LFG-parse of two parallel sentences, we are able to make precise alignments on both the c-structure and f-structure level.

Although there exists many methods for automatic phrase alignment, e.g. [3], most of these have been based on aligning any N-gram that is compatible with a word alignment, where none of these take into account syntactic features, and alignments may cross constituent borders. [2] describe a method for using statistical word-alignments as seeds to two separate constituent and dependency tree alignments; however, the goal here is to create a set of N-gram pairs for statistical machine translation, and the dependency and constituent alignments do not inform each other.

Our method is instead based on the fact that similar grammatical phenomena in different languages will have similar grammatical analyses, so structural similarity in the analyses should indicate that those parts of the analyses may be linked. How much structural similarity is required in order to link two elements is defined as a set of general constraints. This allows for a more top-down method of phrase alignment, which is more informative to the linguist since it links not only true constituents, but functional elements (which in LFG may even span discontiguous constituents). Word-alignments or translational dictionaries may be needed to automatically disambiguate in cases where the LFG parses do not give sufficient information; but the method will perform a large part of the alignment job even without *any* parallel corpus available.

The principles and constraints for alignment are discussed in the next section, section 3 describes the implementation, while section 4 discusses the strengths and weaknesses of the method.

2 Principles for Phrase Alignment

We want our alignment links to be useful for treebank studies, in the XPar-project this includes studying the relationship between syntactic function and semantic roles across languages, thus the principles for alignment (or, constraints on possible alignments) have to take this goal into account. An outline of the principles for phrase alignment used in the XPar-project are formulated in [1, pp. 75–77], this section recounts the major points, and explains some relevant LFG-terminology and concepts.

To introduce the relevant LFG-terminology, consider figure 1. This shows two simplified LFG f-structures and c-structures, ready for alignment. The English word *slept* is a verb phrase, and its nodes *project* the f-structure g (as seen by the PRED value being the 'semantic form' of *slept*, '**sleep**'). The projection from c-structure to f-structure, ϕ, is a many-to-one mapping, and all the nodes S, VP and V together project g. Since the nodes project the same f-structure, they constitute a *functional domain*. We can see that they project the same f-structure by the $\uparrow\ =\downarrow$ annotations, which are read as "my f-structure is the same as that of my mother node". The NP node has $\uparrow$ SUBJ $=\downarrow$ instead, read as "my f-structure is the SUBJ of my mother's f-structure"; the NP thus projects the value of the SUBJ f-structure inside g.

The argument structures of the Norwegian and English verbs are shown in their PRED values; both verbs take one argument, in the figure this is represented by an index. By looking up this index, we find that the one argument of '**sove**' is the subject of f, with '**eg**' as its PRED; similarly '**I**', subject of g, is the only argument of '**sleep**'. Neither of these subjects take any arguments themselves.

The candidates we consider for alignment are c-structure phrases, individual words, and f-structures that contain PRED elements[1]. In figure 1, we can link the

[1]We could consider aligning other f-structure elements, but only PRED elements are sure to exist in both languages, while grammatical features such as $\mathrm{ASPECT}_{\{\}}$ might not exist in both languages, or be possible to link in a one-to-one-manner.

PRED elements of f and g; by doing this we consider their f-structures linked. The PRED values of their arguments are also candidates for alignment, and in this case there would be no reason not to link them.

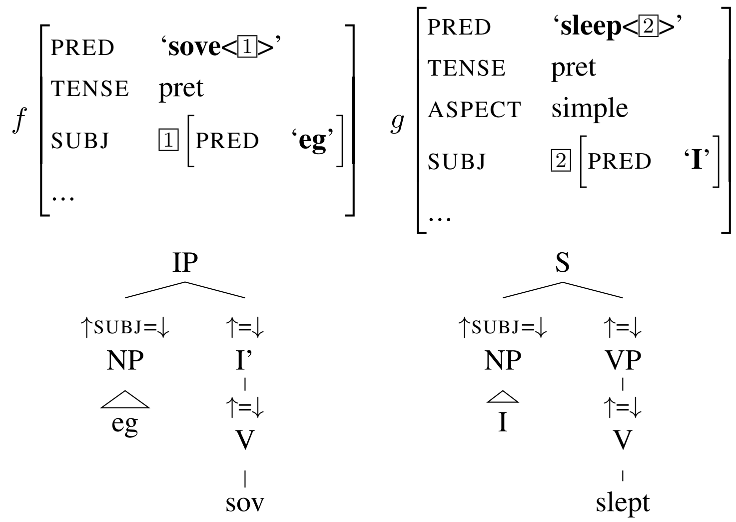

Figure 1: Example of simple links between constituents, f-structures and words (Norwegian and English)

3 Implementation

4 Discussion

5 Conclusion

References

[1] Helge Dyvik, Paul Meurer, Victoria Rosén, and Koenraad De Smedt. Linguistically motivated parallel parsebanks. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, and Frank Van Eynde, editors, *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, pages 71–82, Milan, Italy, 2009. EDUCatt.

[2] M. Hearne, S. Ozdowska, and J. Tinsley. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. In *Actes de la 15e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN '08)*, Avignon, France, 2008.

[3] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.