

Parallel LFG Grammars on Parallel Corpora: a Base for Practical Triangulation

Over many years, the LFG ParGram project (Butt et al., 2002) has demonstrated many advantages of a carefully controlled common framework for grammar development in different languages, not just by picking a common theoretical framework and practical development platform, but by active exchange about representational and design decisions in the grammars. Well-known advantages include the easy cross-lingual migration of modules or applications building on the parser output (f-structures) and the speed-up in building a new grammar when there are existing grammars with worked-out analyses for phenomena shared by the new language. Kim et al. (2003) even demonstrated that, due to the structural similarity between Japanese and Korean, a grammar for Korean could be created in a very short time by starting from the Japanese grammar, replacing the lexicon and then modifying the grammar rules as needed.

We present an architecture for cross-lingual linguistic and language-technological research that is inspired by these known advantages of parallelism (“by design”) across grammars, and combines them with (i) the Annotation Projection idea from recent data-driven work in Natural Language Processing (NLP), exploiting parallelism (“by nature”) in available translation data (Yarowsky et al., 2001), and (ii) the long-standing idea of using multi-parallel text (three or more languages in translational correspondence) for better informed data-driven inference (the Triangulation idea formulated by Martin Kay in a Machine Translation context).

The main idea is very simple: given a (relatively large) multi-parallel corpus and at least two parallel broad-coverage grammars (say, for English and German), the corpus is preprocessed by standard sentence alignment and statistical word alignment routines. The existing grammars are used to parse the versions of the corpus in the respective languages. Word alignment is used to link the words in the “known” languages to words in the “unknown” language(s) in the multi-parallel corpus (in our example Dutch).

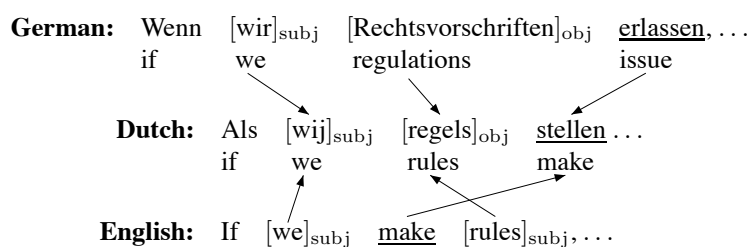


Figure 1: Projection across parallel corpora.

As is shown in Figure 1, the word links can now be used to “project” (f-)structural information, or at least dependency information, from English and German to Dutch.¹ By using two (or more) “known” languages (with parallel grammars) as base for projection, chances are increased that systematic or incidental divergencies are detected by non-parallelism in the “known” languages’ analyses (as shown in Figure 2). Hence, there is a natural way of associating the (partial) projected information with confidence levels (which can then be used for semi-automatic linguistic annotation or for bootstrapping NLP tools).

In practice, the picture is somewhat noisier than our first outline suggested. Real linguistic divergences are not the only reason for observed non-parallelism. The statistical word alignment and the automatic parsing/disambiguation of the “known” languages are of course sources for errors. To make the idea useful in practice, we therefore combine the multi-base projection idea with machine learning techniques that are able to generalize from a number of typical examples. There are two ways of obtaining the typical examples:

¹The words introducing the PRED values in English and German are chosen as the representatives for an entire argument, tracing back the ϕ projection to the PRED value’s c-structural origin. Our focus here is not the full phrasal span of argument phrases, but their lexical exponents.

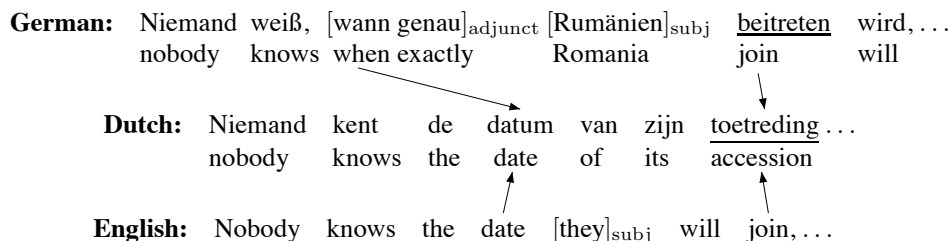


Figure 2: Cross-language divergencies.

either automatically, by using very strict confidence level criteria in the projection scenario; or by adopting an Active Learning strategy, i.e., by selecting the most informative instances for a human annotator to hand-label.

In this contribution, we present an architecture for performing multi-base projection experiments with a number of different strategies, including a combination of individual (“local”) argument classification decisions in a global ranking approach. Our current results are based on the Europarl (Koehn, 2005) corpus, focusing on the three mentioned languages for method development.² For machine learning, we applied the MegaM software package.³ Our preliminary “local” argument classifier (based on a training set of just a few hundred sentence triples, from which the high-confidence examples were chosen) has a precision of 69.8% and recall of 36.3% (F-score: 47.8%) when trained on automatically selected training instances (exploiting parallelism); with simulated manual annotation (using the Alpino parser output as the training data annotation), a precision of 79.5% and recall of 51.0% is reached (F-score: 62.2%).⁴ Large-scale experiments, detailed analyses and global ranking experiments are currently under way.

The proposed framework is very flexible and open to extensions in a variety of directions. One of the immediate goals of our research is to support linguistic exploration of large, unannotated corpora in linguistic research on phenomena that do not occur with high frequency (in particular, interaction of factors influencing information structure).

References

- Bouma, G., van Noord, G., and Malouf, R. (2001). Wide coverage computational analysis of dutch. In Daelemans, W., Sima’an, K., Veenstra, J., and Zavrel, J., editors, *Computational Linguistics in the Netherlands, CLIN 2000*, pages 45–59, Amsterdam. Rodopi.
- Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.
- Kim, R., Dalrymple, M., Kaplan, R. M., and King, T. H. (2003). Porting grammars between typologically similar languages: Japanese to korean. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC-17)*.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit 2005*.
- Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001*.

²By focusing on Dutch as the target “unknown” language, we able to use a reliable parser outputting dependency structures for comparison: the Alpino parser (Bouma et al., 2001).

³<http://www.cs.utah.edu/~hal/megam/>

⁴The current evaluation results are taken from a development set of unseen sentence triples, for which the Alpino output was taken as the gold standard. A manually annotated test set will not be used until the experimental development has been refined. The Alpino parser has a precision of 86.6% and recall of 90.6% (F-score: 88.5%) against our manually labelled test set.