Using parallel corpora to bootstrap multilingual semantic parsers

Kilian Evang
University of Groningen
k.evang@rug.nl

Johan Bos
University of Groningen
johan.bos@rug.nl

A semantic parser is a system that maps natural-language utterances to machine-interpretable meaning representations. Semantic parsers with wide coverage are available nowadays [1]. But they are usually made for a single natural language (English, in most cases), and to develop one for another language is time-consuming and costly. Given a semantic parser for English and a parallel corpus, we propose a new method to learn a semantic parser for a different language.

To date, most work on learning semantic parsers from text/meaning pairs has been done on very limited natural-language domains. Due to an explosion in search space, the step to larger domains is a considerable challenge that has first been tackled only recently [3].

The basic method we propose is as follows: the English portion of a sentence-aligned parallel corpus is annotated using the C&C parser for Combinatory Categorial Grammar [4, CCG] and the Boxer semantic construction system. This produces not only sentence/meaning pairs, but also CCG derivations as in Figure 1. The meaning representations follow Discourse Representation Theory (DRT) [2]; we abbreviate them here with simple symbols. From these derivations, we can read off an inventory of lexical categories for English, i.e. syntactic-semantic categories of individual words.

We hypothesize that every lexical category needed for parsing the target language (say, Dutch) can be generated from the English inventory by one or more of the following operations: (i) copying an English category, (ii) reversing the directionality of the syntactic category, (iii) combining two lexical categories into one using combinatory rules such as application or composition, (iv) splitting a category into two, as if generated by a combinatory rule. For example, Figure 1 shows a combined English and Dutch derivation for one sentence pair. The Dutch words *john* and *zwemt* have categories that are identical to their English counterparts. The word *graag*, however, had its category built by composing those of *likes* and *to* and reversing the directionality of the category.

Guided by this hypothesis, the task of the learning component is to 1) build the right category inventory for the target language and associating target language words with the right categories, yielding a target language *lexicon*, 2) learn a statistical parsing model for the target language.

We are currently approaching this task with a parallelized version of a perceptron learner as in Zettlemoyer and Collins [6]. It works by repeatedly generating candidate lexical items and according derivations for a sentence, testing whether the generated semantics is the desired one, and accordingly weighting up or down the lexical items involved in the parses.

The project has just started. We are currently working on first experiments with short sentences from the EMEA parallel corpus [5] and target languages similar to English, *viz.* Dutch and German. The central challenge is to develop an effective way to constrain the search space over possible generated categories for the target language, since an exhaustive search is impossible. We then aim

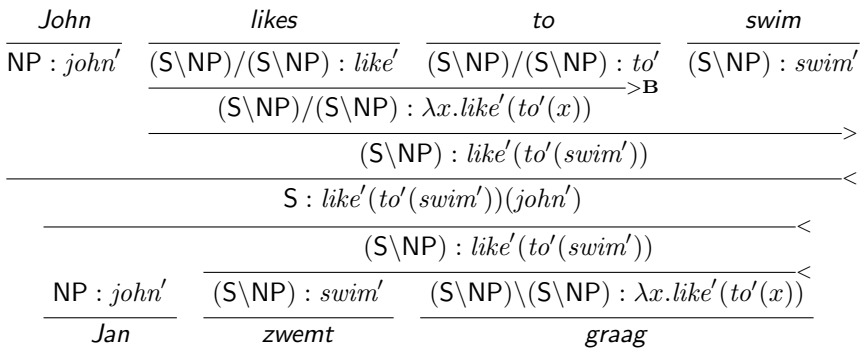

Figure 1: Paired CCG derivations of English (downwards) and Dutch (upwards) sentence with identical semantics.

to move towards longer sentences or even texts, and to target languages typologically more distinct from English.

References

[1] James Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale nlp with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, 2007.

[2] Hans Kamp. A Theory of Truth and Semantic Representation. In Jeroen Groenendijk, Theo M.V. Janssen, and Martin Stokhof, editors, *Truth, Interpretation and Information*, pages 1–41. FORIS, 1984.

[3] Phong Le and Willem Zuidema. Learning compositional semantics for open domain semantic parsing. In *Proceedings of COLING 2012*, pages 1535–1552, 2012.

[4] Mark Steedman. *The Syntactic Process*. The MIT Press, 2001.

[5] Jörg Tiedemann. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, 2009.

[6] Luke Zettlemoyer and Michael Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, 2007.