Using parallel corpora to bootstrap multilingual semantic parsers

Kilian Evang
University of Groningen

Johan Bos
University of Groningen

Problem

Existing wide-coverage deep semantic parsers are language-specific (usually English)

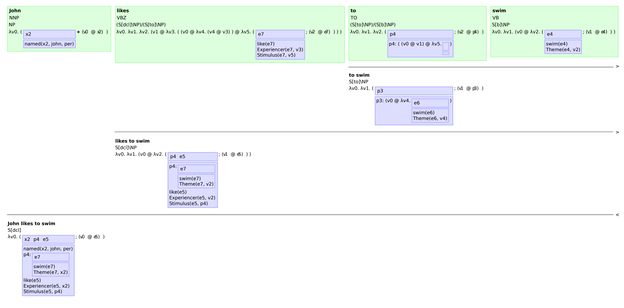
Can we exploit parallel corpora to transform a semantic parser for English into a semantic parser for another language?

Lexicalized grammar

Starting point: English derivations produced by C&C parser and Boxer

CCG syntax, DRT formal semantics

For a new language, we need to learn:
- a new lexicon
- new parser weights



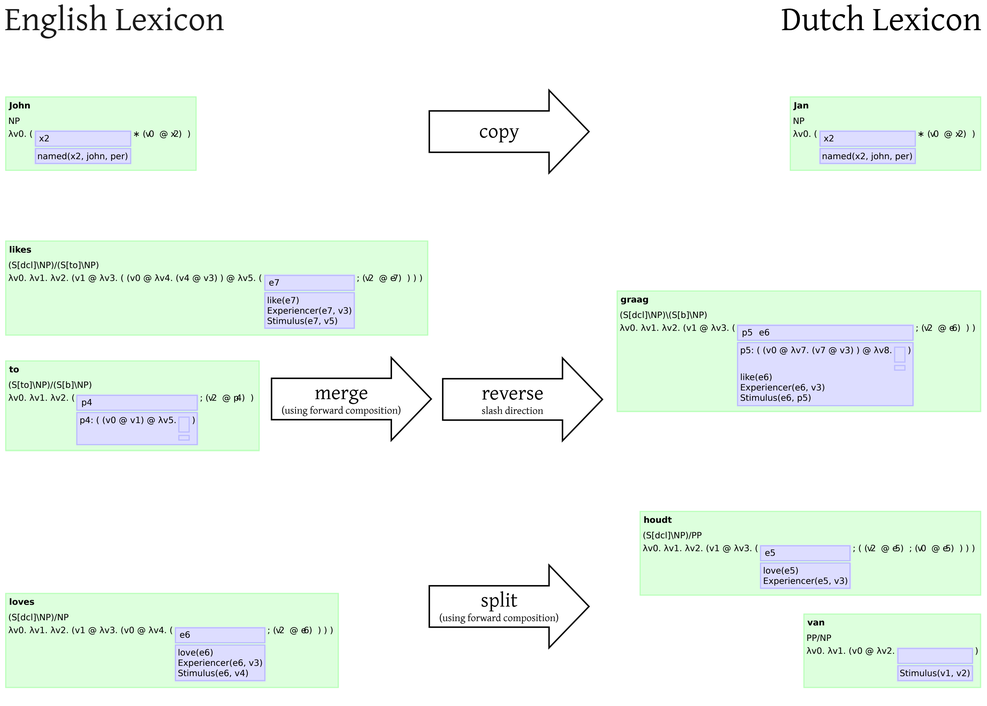
Transforming the lexicon

Consider the following English/Dutch parallel sentences:

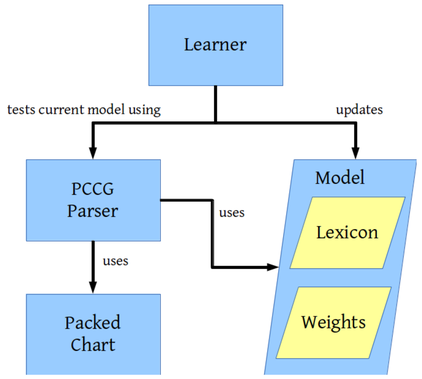
(1a) John likes to swim.
(1b) Jan zwemt graag.

(2a) John loves Mary.
(2b) Jan houdt van Maria.

We hypothesize that these four operations are enough to transform an English lexicon into a lexicon for another language:



Learning



Perceptron learning algorithm following Zettlemoyer and Collins (2007) learns lexicon and weights

Features: lexical entries, use of unary CCG rules, more coming soon

Results so far

Working with EMEA corpus (European Medicines Agency), think drug package inserts

First experiments with restricted setting: short English sentences

23% sentence-level accuracy
(relation-level higher, but harder to measure)

Promising, but much remains to be done!

Goals

Deal with longer sentences: use chart pruning and/or supertagging

Deal with names, numbers and other unseen words: use semantic templates

Make the move to different languages: effectively constrain the search space for lexicon transformation