Case and Content: A Cross-Linguistic Corpus Study

Drew Reisinger (reisinger@cogsci.jhu.edu) · Johns Hopkins University

Summary. I present two projects designed to enable cross-linguistic data-driven approaches to studying argument realization and morphological case frames. First, I investigate cross-linguistic projection techniques that leverage alignment tools from machine translation to construct new semantically annotated corpora in another language (in this case, Czech) from existing resources in English. Second, I present a preliminary model of how a verb's lexical semantics contribute to its arguments' syntactic expression and morphological case marking, a problem which I call *morphosyntactic argument realization*. Among my findings from these investigations are (1) that alignment models from machine translation introduce too much noise into predicate and argument alignments to be useful for linguistic study and (2) that models of morphosyntatic argument realization must be able to capture the heavily skewed case distributions that appear in naturalistic corpora.

Background. In reaction to the problems of role specification and fragmentation facing thematic role theories of argument realization, Dowty [1991] proposes *thematic proto-roles* as an alternative representation of thematic content. Instead of associating each verbal argument with one of a possibly large set of imprecise categorical roles, he describes the argument in terms of which of a small set of privileged entailments, which I will call *proto-role properties*, it satifies. A verb's syntactic expression is then a function of the proto-role properties of its arguments, and traditional thematic roles emerge as fuzzy clusters of sets of entailments analogous to prototype concepts.

Subsequent work has validated Dowty's approach on large-scale datasets. In partiuclar, Reisinger et al. [2015] construct a crowdsourced corpus of proto-role property annotations on a subset of the Proposition Bank corpus [Palmer et al., 2005] in support of a new NLP task, *semantic proto-role labeling* (SPRL). The annotation protocol consists of answering a series of "How likely" questions on a five-point Likert scale and can be completed by annotators with relatively little training, such as those recruited on Amazon Mechanical Turk. White et al. [2016, in review] then show that a probabilistic implementation of Dowty's proto-role linking theory predicts subject selection well on this SPRL corpus.

Separately, Grimm [2005, 2011] extends Dowty's theory in a different direction by arguing that morphological case distributions in a variety of languages can be explained in the same framework as argument realization. In particular, he claims that morphological case, like syntactic expression, is a function of an argument's proto-role properties. The projects I present are a first step toward a wide-coverage empirical validation of Grimm's proposal

Methods. To automatically construct a corpus of Czech verbs and arguments annotated with proto-role properties, I project the SPRL annotations from the Reisinger et al. [2015] corpus to its Czech translation provided by the the Prague Czech-English Dependency Treebank [Hajič et al., 2012], a manually translated parallel corpus with morphological annotations (among other kinds), using the Berkeley Aligner[1] to identify Czech verbs and arguments that correspond with each SPRL-annotated English verb and argument. Because this alignment step is noisy, I then apply several layers of filtering heuristics, such as requiring English verbs to align with Czech verbs, to remove alignments that are likely to be incorrect.

In order to evaluate how well the projected SPRL judgments can be used to predict morphological case on the Czech dataset, I propose a model based on SVM^{rank} [Joachims, 2006][2] which ranks possible assignments of cases to arguments based on the arguments' proto-role entailments. This model is trained on the previously described projected Czech corpus as well as on the English

[1] `https://code.google.com/archive/p/berkeleyaligner`

[2] `https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

Universal Dependencies (UD) corpus [Nivre et al., 2015] annotated for SPRL by White et al. [2016, in review].

Findings. Even though the heuristic filters used to remove potentially incorrect alignments in the projection process are relatively conservative, I was unable to project SPRL annotations from a large number of verb-argument pairs. Furthermore, many of the remaining alignments are still incorrect, and this alignment noise contributes to the errors made by the case prediction model. Thus, it seems that automatic alignment techniques used for machine translation are not ideal for the specific task of projecting predicate-argument annotations.

Despite the case-prediction model's simplicity, it performs well at predicting argument realization on the UD English corpus, although the skewed distribution of syntactic configurations in naturalistic data precludes evaluating the model's performance on ditransitives. However, it performs relatively poorly at predicting Czech case on the projected corpus, reflecting (1) the heavily skewed distribution of nominative and accusative cases over more oblique cases, (2) the inadequacy of the model for phenomena such as null subjects and valency changes, and (3) the significant noise introduced by the automatic alignment step.

References

D. Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

S. Grimm. The lattice of case and agentivity. Master's thesis, University of Amsterdam, 2005.

S. Grimm. Semantics of case. *Morphology*, 21:515–544, 2011.

J. Hajič et al. Prague Czech-English Dependency Treebank 2.0. LDC2012T08. Linguistic Data Consortium, 2012.

T. Joachims. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

J. Nivre et al. Universal dependencies 1.2, 2015. URL `http://hdl.handle.net/11234/1-1548`. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31:71–106, 2005.

D. Reisinger, R. Rudinger, F. Ferraro, C. Harman, K. Rawlins, and B. V. Durme. Semantic proto-roles. *Transactions of the ACL*, 3:475–488, 2015.

A. S. White, D. Reisinger, R. Rudinger, K. Rawlins, and B. V. Durme. Computational linking theory. *Transactions of the ACL*, 2016, in review.