Unsupervised discovery of Construction Grammar representations from unannotated corpus for under-resourced languages

Bogdan Babych
University of Leeds

Computational Linguistics applications and corpus annotation tools rely on morphological and syntactic resources for different languages: part-of-speech taggers, lemmatizers, dependency or constituency parsers. Creating these resources for a new language usually involves two separate stages: a grammar development stage (i.e., creating or adjusting a tagsets for language-specific morphosyntactic features, manually disambiguating tags in a sub-corpus used for training a tagger, manually checking a training treebank, etc.) and a lexicon-development stage (i.e., identifying possible emission tags for word forms, paradigm classes for lemmas and lemmatization operations for tagged word forms). Typically this development is done in a theory-neutral way (e.g., Straka & Straková, 2017), which often means that the annotation scheme contains linguistically unsound or contradictory solutions that lead to potential errors and reduce usefulness of the annotation.

Our paper describes an on-going project on systematic linguistic development of lexicogrammatic corpus annotation tools for under-resourced languages in a lexicalized framework of Construction Grammar (Kay & Fillmore, 1999; Fillmore, 2002), which integrates lexicosemantic and morphosyntactic representations in a coherent theoretical framework and involves a single-stage induction of morphosyntactic lexicon of single- and multiword expressions enriched with information about their morphological variation, valencies, subcategorization frames, collocations, semantic prosodies and phraseological expressions. These representations are formalized as an extension to probabilistic Tree Adjoining Grammar – TAG (Kasai et al., 2017; Bangalore & Joshi, 1999) with the syntactic groups formalism (Gladkii, 1985). Even though both Construction Grammar theory and TAG parsing are active areas of research, there has been a gap in joining these areas of modeling and processing within a linguistically motivated corpus-based framework.

The central challenge for our project is that traditional approaches to grammar induction and lexicon development often rely on supervised or semi-supervised methods using manually annotated or checked corpora, which are not available for many smaller and less-resourced languages. We present an alternative methodology of unsupervised Construction Grammar induction from unannotated corpora, which involves (1) manual specification of inflection tables and patterns of possible phonological alternations in paradigms using linguistic descriptions (e.g., a published grammar books) for a given under-resourced language; (2) automated induction of the morphological lexicon by segmenting word forms using tables of affixes from the inflection tables, verifying hypotheses about paradigms with the corpus data via concatenating a candidate stem with other inflections in each matching inflection table and applying matching morphological distortion patterns; (3) deriving string representations for candidate constructions using corpus-based association measures (log-likelihood and mutual information) between word forms, lemmas and combinations of morphological features for sets of morphologically annotated skip-grams (N-grams with additions of possible gaps) in the corpus; (4) identifying syntactic relations and lexicalized TAG representations for candidate constructions using data-oriented parsing techniques (Beekhuizen & Bod, 2014).

We report evaluation results for a large-scale construction grammar induction for Ukrainian based on a 200 million words news corpus, that currently covers around 50k single and multiword constructions. Compared to existing approaches to lexicon induction, such as (Ahlberg et al, 2015), the proposed approach does not require seed annotated data for

supervised training, does not miss any inflection types, it is more robust against accidental noise in corpus, since it performs a latent, indirect induction of paradigms, e.g., by allowing for the same word form to provide evidence and support hypotheses for several paradigms, allowing for morphological ambiguity.

References

Ahlberg, M., Forsberg, M., & Hulden, M. (2015). Paradigm classification in supervised learning of morphology. In Proc. of NAACL

Bangalore, S. and Joshi, A.K., 1999. Supertagging: An approach to almost parsing. Computational linguistics, 25(2), pp.237-265.

Beekhuizen, B. and Bod, R. 2014. Automating construction work: Data-Oriented Parsing and constructivist accounts of language acquisition Automating Construction Work. In: Data-Oriented Parsing and Constructivist Accounts of Language Acquisition. Extending the Scope of Construction Grammar, Mouton, Berlin, pp.47-74.

Fillmore, C.J. 2002. Mini-grammars of some time–when expressions in English. In: Bybee, J. & Noonan, M. (eds) Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson. Amsterdam & Philadelphia: Benjamins. 31–59.

Gladkii, A.V., 1985. Syntactic Structures of Natural Language in Computer-Aided Communication Systems. Nauka, Moscow.

Kasai, J., Frank, B., McCoy, T., Rambow, O. and Nasr, A., 2017. Tag parsing with neural networks and vector representations of supertags. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1712-1722).

Kay, P. and Fillmore, C.J. 1999. Grammatical constructions and linguistic generalizations: What's X doing Y? construction. Language 73 1: 1–33.

Straka, M., Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.