RELATIONAL-REALIZATIONAL PARSING

REUT TSARFATY
UPPSALA UNIVERSITY

Statistical parsing models aim to assign accurate syntactic analyses to natural language sentences based on the patterns and frequencies observed in human-annotated training data. State-of-the-art statistical parsers to date demonstrate excellent performance in parsing English, but when the same models are applied to languages different than English, they hardly ever obtain comparable results. In this talk I present a new parsing model, called the *Relational-Realizational (RR)* model [11], that is designed to effectively cope with parsing languages that allow for flexible word-order and richer morphological marking. The model is developed based on two principles: the first is *form-function separation* in the representation of constituents, and the second is *typological decomposition* of the syntactic spell-out rules. I show an application of the RR model to parsing the Semitic language Modern Hebrew, obtaining significant improvements over previously reported results.

Many state-of-the-art statistical parsing models to date have been developed with English data in mind, utilizing the Wall-Street Journal Penn treebank [5] as the primary, and often the sole, resource. However, English is quite unusual in its fairly *configurational* character [2]. The main challenge associated with parsing languages that are *less configurational* than English, such as German, Arabic, Hebrew or Warlpiri, is the need to model and to statistically learn complex correspondence patterns between functions, e.g., sets of abstract grammatical *relations*, and their morphological and syntactic forms of *realization*.

Whereas grammatical relations are largely universal, realization is known to vary greatly. Different means of realization involve the interaction of (at least) two typological parameters, one associated with word order [4], and another associated with word-level morphology [8, 3]. In order to adequately model complex form-function correspondence patterns that emerge from such interactions, I firstly consider morphological models that map grammatical properties of words to the surface formatives that realize them, and I follow up on recent studies in singling out the paradigmatic, realizational approach as an adequate strategy for modeling complex form-function correspondence [1]. I then extend the *inferential, realizational* principles of mapping grammatical properties to surface words [10] to *relational, realizational* principles of mapping grammatical relations to surface constituents.

In the resulting RR model, constituents are organized into syntactic paradigms [6]. Each cell in a paradigm separates the function, a Relational Network [7] and a set of grammatical properties, from the form of the individual constituent. The form of a constituent in a specific paradigm cell emerges from the (i) internal grouping, (ii) linear ordering, and (iii) morphological marking of its subconstituents. The RR decomposition of spell out rules for specific paradigm cells into parameters thus separates the *functional, configurational* and *morphological* dimensions. Subconstituents may belong to different paradigms and may

be associated with relational networks of their own, and the process continues recursively until fully-specified morphosyntactic representations map to words. This 3-phased spell-out gives rise to a recursive generative process that can be used as a probabilistic model and its probabilistic parameters can be estimated from data based on relative frequencies.

The resulting statistical model is empirically evaluated by parsing sentences in the Semitic language Modern Hebrew on the basis of a small annotated treebank [9]. Through a series of experiments we report significant improvements over the state-of-the-art Head-Driven (HD) alternative on various measures, without paying any computational costs. The typological characterization of the RR statistical distributions also suggests itself as a starting point for the development of quantitative methods that would facilitate typological classification learned directly from natural language corpus data.

Acknowledgements

The research reported in this talk constitutes a part of a 4-year project entitled "Integrated Morphological and Syntactic Ambiguity Resolution for Modern Hebrew" (principal invetigator: Reut Tsarfaty) funded by the Dutch Science Foundation (NWO), grant number 017.001.271. This work was supervised by Dr. Khalil Sima'an and Prof. Remko Scha.

References

[1] Stephen R. Anderson. *A-Morphous Morphology*. Cambridge University Press, 1992.
[2] Joan Bresnan. *Lexical-Functional Syntax*. Blackwell, 2000.
[3] Joseph H. Greenberg. A quantitative approach to the morphological typology of language. In R. F. Spencer, editor, *Method and Perspective in Anthropology*. University of Minessota Press, Minneapolis, 1954.
[4] Joseph H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, 1963.
[5] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.
[6] Kenneth L. Pike. Dimensions of grammatical constructions. *Language*, 38(3):221–244, 1962.
[7] Paul M. Postal and David M. Perlmutter. Toward a universal characterization of passivization. In *BLS 3*, 1977.
[8] Edward Sapir. *Language: An Introduction to the Study of Speech*. Brace and company, New York, 1921.
[9] Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. Building a Tree-Bank for Modern Hebrew Text. In *Traitement Automatique des Langues*, 2001.
[10] Gregory T. Stump. *Inflectional Morphology: A Theory of Paradigm Structure*. Number 93 in Cambridge Studies in Linguistics. Cambridge University Press, 2001.
[11] Reut Tsarfaty. *Relational-Realizational Parsing*. PhD thesis, University of Amsterdam, 2010.