A Transition-Based Framework for Morphological, Syntactic and Joint Morphosyntactic Parsing of Morphologically Rich Languages (MRLs)

Amir More
School of Computer Science
The Interdisciplinary Center
Herzeliya, Israel
more.amir@idc.ac.il

Reut Tsarfaty
Mathematics and Computer Science
Weizmann Institute of Science
Rehovot, Israel
reut.tsarfaty@weizmann.ac.il

Abstract

State-of-the-art results for morphosyntactic analysis of Morphologically Rich Languages (MRLs) such as Hebrew are currently too low to enable real-world applications that are successfully implemented for heavily-studied languages like English. This is because existing frameworks for (morpho)syntactic analysis rely on fundamental structuralists' assumptions, and in particular, assume a strict separation between morphological and syntactic processing. This assumption breaks down in the context of MRLs. In this work we present a general-purpose shift-reduce framework that implements a standalone morphological analyzer, a standalone dependency parser, and a joint morphosyntactic dependency parser for MRLs. Each of these tasks is defined via a transition system, a cross-linguistic feature model, and a global scoring function. The learning problem is solved using the structured perceptron, and efficient decoding is achieved via beam-search. We present state-of-the-art results for each of the tasks in isolation, and present the first joint transition-based system for morphological segmentation and dependency parsing.

Background Dependency parsing has seen a surge of interest in recent years, with dependency trees amenable for intuitive semantic interpretation and efficient processing (Kübler et al., 2009). State-of-the-art dependency parsing results reported for English present high accuracy on Labeled Attachment Score (LAS) (McDonald et al., 2005; Nivre, 2007b; Zhang and Nivre, 2011). The model of Zhang and Nivre (2011), for instance, achieves LAS of 91.8% on the standard English benchmark, high enough to enable a variety of applications, such as machine translation and information extraction. For many other languages, state-of-the-art results are significantly lower. This is particularly so for languages called Morphologically Rich Languages (MRLs), as shown in various shared tasks (Buchholz and Marsi, 2006; Nivre, 2007a; Seddah et al., 2013). This work addresses the challenge of developing a dependency parser that can effectively cope with morphological, syntactic, and morphosyntactic phenomena in MRLs.

The Challenge A common practice in dependency parsing is to assume a strict separation between morphological and syntactic processing. This effectively means that a morphological disambiguation component initially assigns POS tags and morphological features to surface words, and these morphologically annotated words are the input for the syntactic parser. Moreover, in order to parse MRLs, morphological disambiguation is used not only for assigning POS tags and morphological features to words, but also for translating input tokens to their constituent morphemes, each of which carries its own POS tag and features. In the context of Modern Hebrew, this pipeline approach is attempted in Goldberg and Elhadad (2010). However, their results indicate that this separation significantly impacts parsing performance. While adequate results are achieved for correctly disambiguated data (gold-standard morphology), parsing performance using automatically disambiguated morphemes is significantly lower.

Method/Approach We hypothesize that applying joint morphological and syntactic analysis to MRLs provides an opportunity for a rich, cross-linguistic model, enabling access to partial syntactic analysis for morphological disambiguation – and vice-versa. This approach has been successfully employed in phrase-structure parsing for MRLs (Cohen and Smith, 2007; Goldberg and Tsarfaty, 2008; Green and Manning, 2010; Goldberg and Elhadad, 2011), but dependency parsers for MRLs still assume a strict pipeline (Seddah et al., 2013). On top of that, statistical parsing models are often customized per language, where the weights of features from pre-defined feature-templates are automatically induced

from corpus data. Developing a cross-linguistic feature model – that is, a language independent feature set that performs equally well on different languages – would enable a one-size-fits-all solution, but the dominance of properties of heavily-studied languages prove to be a formidable barrier. We conjecture that for such cross-linguistic portability to be achieved, the addition of new dimensions to the linguistic feature model is required, and in particular, adding dimensions that cover properties of MRLs.

The Solution We present a general-purpose transition-based parsing suite, with which we instantiate different statistical models for different MRL processing tasks. We present a model for morphological disambiguation, for dependency parsing, and for joint morphosyntactic dependency parsing. Each of the standalone models is based on its own transition system, while the joint model unifies actions from both levels of processing. Each of the models is learned based on a global scoring function, taking into account the transition sequences of the relevant system, and a feature model. Here too, the joint system unifies features from both levels of processing. The learning procedure is based on the structured perceptron, which is optimized for complete morpho-syntactic transition sequences. The search for the best parse is based on beam-search decoding, which considerably increases the space of parsing hypotheses compared to greedy parsers, retaining parsing efficiency.

Experiments We used our framework to instantiate a dependency model that reproduces the state-of-the-art dependency parsing results by Zhang and Nivre (2011) for English. This parser is based on the Arc-Eager transition system (Nivre, 2007b) and a set of rich non-local features indexed according to the position and direction of dependents relative to the head (Zhang and Nivre, 2011). We then introduced an Arc-Standard transition system with a cross-linguistic feature model that is word-order agnostic, and which relies on rich morphological information, while maintaining results comparable to state-of-the-art. We further introduced a new transition-based system for morphological disambiguation, based on the same general-purpose framework, where we design a novel semi-open transition system whose actions choose between morpheme options of a node in an ambiguous lattice, such that a the disambiguated path is constrained to occur within the lattice. For the feature model, we use feature templates of uni-gram, bi-gram, and tri-gram of morphemes. Building on these two models, we test our hypothesis concerning joint processing via a unified transition system, allowing both morphological and syntactic transition actions, and taking into account the two kinds of features. We test various strategies for alternation between morphological and syntactic transitions and beam-alignment solutions.

Results Our dependency parsing results match Zhang and Nivres state-of-the-art results for English, with many parses indistinguishable from theirs. Applying this model to Hebrew achieves state-of-the-art results of UAS 85.92% and LAS 76.59%. Applying our word-order agnostic cross-linguistic model, we obtained state-of-the-art results of UAS 86.43% and LAS 76.95%, compared to UAS 84.4% of Goldberg and Elhadad (2010). For our standalone morphological model we obtained F1 scores of 90.73% for segmentation including morphological properties, 91.5% for POS tagging and segmentation and 96.44% for only segmentation on the Hebrew standard split (in comparison to 88.5% and 92.32% for full morphological disambiguation and POS tagging, respectively (Adler and Elhadad, 2006)).

Discussion The immediate concern for our joint parser is the alignment of candidates in the beam. Since different disambiguation decisions can result in varying sentence lengths, candidates in the beam may have a different number of transitions. This results in two types of bias: preferring short-sequence candidates that terminate early with the highest score when they terminate (i.e. not giving a chance for longer-sequence candidate), and preferring long-sequence candidates which have more features. Alignment can mitigate this by making sure all candidates advance and compete in lock-step, such that short sequences are required to wait for longer ones. Relative scoring can add weights to short sequences such that they maintain relevance when compared to longer sequences. We are currently experimenting with various state-alignment strategies and weight along these lines, for improving our joint parser.

Conclusion We introduce a transition-based standalone morphological disambiguator for Modern Hebrew, a cross-linguistic word-order agnostic adaptation of Zhang and Nivres dependency parser, and unify the two using a novel joint morphosyntactic processor. Independently, our components obtain state-of-the-art results. The work on our joint processor is underway, experimenting with different alignment strategies that would allow us to test our joint processing hypothesis.

References

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based hmm for hebrew morphological disambiguation. In *Proceedings of COLING-ACL*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL-X*, pages 149–164.

Shay B. Cohen and Noah A. Smith. 2007. Joint morphological and syntactic disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 208–217.

Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Proceedings of ACL*.

Yoav Goldberg and Michael Elhadad. 2011. Joint Hebrew segmentation and parsing using a PCFGLA lattice parser. In *Proceedings of ACL*.

Yoav Goldberg and Reut Tsarfaty. 2008. A single framework for joint morphological segmentation and syntactic parsing. In *Proceedings of ACL*.

Spence Green and Christopher D. Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of COLING*.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Number 2 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.

Joakim Nivre. 2007a. The conll 2007 shared task on dependency parsing.

Joakim Nivre. 2007b. *Inductive Dependency Parsing*. Springer.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam PrzepiÓrkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolisk, Alina WrÓblewska, and Éric Villemonte De La Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, United States. Association for Computational Linguistics.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of ACL*.