A Transition-Based System for Joint Lexical and Syntactic Analysis (WG3)

Matthieu Constant
Université Paris-Est, LIGM
Alpage, INRIA, Université Paris Diderot
France
Matthieu.Constant@u-pem.fr

Joakim Nivre
Uppsala University
Dept. of Linguistics and Philology
Uppsala, Sweden
joakim.nivre@lingfil.uu.se

1 Introduction

The integration of Multiword Expressions (MWEs) in linguistic analysis is a great challenge for Natural Language Processing. From a semantic point of view, there exists a continuum between entirely non-compositional expressions (*piece of cake*) and almost free expressions (*traffic light*). Many MWEs are indeed semi-compositional (*white wine*). MWEs may be decomposable and partially analyzable, which implies the need for predicting their internal structure in order to compute their meaning. From a syntactic point of view, MWEs often have a regular structure and do not need special syntactic annotation. Some MWEs have an irregular structure, such as *by and large* which on the surface is a coordination of a preposition and an adjective. They are syntactically as well as semantically non-compositional and cannot be represented with standard syntactic structures. MWEs can therefore be divided between regular and irregular ones. This dichotomy can be extended to fixed and non-fixed MWEs in the sense of Sag et al. (2002).

This poster[1] presents a novel representation that allows both fixed and non-fixed MWEs to be adequately represented without compromising the syntactic representation. We then show how this representation can be processed using a transition-based system that is a mild extension of a standard dependency parser. This system takes as input a sentence consisting of a sequence of tokens and jointly predicts its syntactic dependency structure as well as its lexical units (including MWEs). It practically shows great improvements in terms of MWE identification as compared with the mainstream joint systems like Candito and Constant (2014) using the fixed/non-fixed dichotomy.

[1]The original presentation of this work can be found in Constant and Nivre (2016).

2 Syntactic and Lexical Representation

Our lexical and syntactic representation combines two factorized substructures: (i) a standard tree representing the syntactic dependencies between the lexical elements of the sentence and (ii) a forest of lexical trees including MWEs identified in the sentence.

Each lexical unit – whether a single word or an MWE – is associated with a *lexical node*, which has linguistic attributes such as surface form, lemma, part-of-speech tag and morphological features. Lexical nodes corresponding to MWEs are said to be *non-terminal*, because they have other lexical nodes as children, while lexical nodes corresponding to single words are *terminal* (and do not have any children). Some lexical nodes are also *syntactic nodes*, that is, nodes of the syntactic dependency tree. These nodes are either non-terminal nodes corresponding to irregular MWEs or terminal nodes corresponding to words that do not belong to a fixed MWE. Syntactic nodes are connected into a tree structure by binary, asymmetric dependency relations pointing from a *head* node to a *dependent* node. Figure 1 shows the representation of the sentence *the prime minister made a few good decisions*. It contains three non-terminal lexical nodes: one fixed MWE (*a few*), one contiguous non-fixed

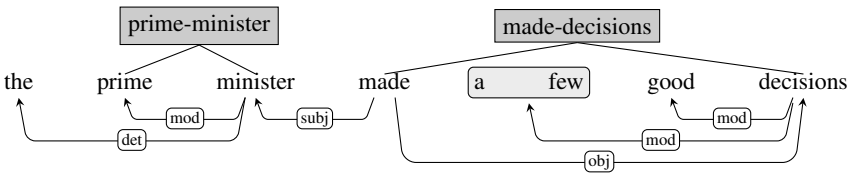

Figure 1: Representation of syntactic and lexical structure.

MWE (*prime minister*) and one discontiguous non-fixed MWE (*made decisions*). Of these, only the first is also a syntactic node. Note that, for reasons of clarity, we have suppressed the lexical children of the fixed MWE in Figure 1. (The non-terminal node corresponding to *a few* has the lexical children *a* and *few*.) Non-fixed MWEs have regular syntax and their components might have some autonomy. For example, in the light verb construction *made-decisions*, the noun *decisions* is modified by the adjective *good* that is not an element of the MWE. The new representation also allows us to represent the hierarchical structure of embedded MWEs.

3 The transition-based model

The proposed parsing model is an extension of a classical arc-standard parser (Nivre, 2004), integrating specific transitions for MWE detection.In order to deal with the two linguistic dimensions separately, it uses two stacks (instead of one): a syntactic stack and a lexical stack. It is synchronized by using a single buffer in order to handle the factorization of the two structures. It includes two types of transitions manipulating such structures: (i) classical arc-standard transitions in charge of adding arcs between lexical nodes, (ii) transitions in charge of creating new lexical nodes (the MWE ones), as well as classifying them (non-fixed vs. fixed ones). The system also includes different hard constraints in order to reduce ambiguities artificially created by the addition of new transitions. A transition sequence in the new system derives the set of lexical nodes and simultaneously builds a projective dependency tree over the set of syntactic nodes. To the best of our knowledge, this system is the first transition-based parser that includes a specific mechanism for handling MWEs in two dimensions.

Experimental results on two datasets (French Treebank and Streusle corpus) show that MWE identification is greatly improved with respect to the mainstream joint approach. In particular, when we used a greedy implementation of our system combined with a simple perceptron model, we observed gains up to two points in F-score for MWE identification.

4 Conclusion

This abstract proposes a transition-based system that extends a classical arc-standard parser for handling both lexical and syntactic analysis. It is based on a new representation having two linguistic layers sharing lexical nodes. This can be a useful starting point for several lines of research such as implementing more advanced transition-based techniques (e.g. beam search, dynamic oracles).

References

Marie Candito and Matthieu Constant. 2014. Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. In *ACL 14 - The 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, United States, June. ACL.

Matthieu Constant and Joakim Nivre. 2016. A Transition-Based System for Joint Lexical and Syntactic Analysis. In *The 54th Annual Meeting of the Association for Computational Linguistics (ACL16)*.

Joakim Nivre. 2004. Incrementality in Deterministic Dependency Parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together (ACL)*, pages 50–57.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.