The Role of Orchestration in Hybrid Parsing of Multiword Expressions

PARSEME WG3 Poster Proposal, January 2015

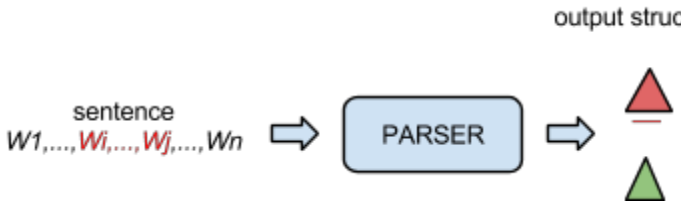
Mike Rosner, Matthieu Constant, Gerold Schneider, Jan Genci, Joakim Nivre

Abstract. The poster described by this proposal is intended to lay the foundations for a discussion of one particular aspect of a forthcoming state-of-the-art report on hybrid parsing and translation of MWEs. That aspect is the activity of parsing itself. In particular, how to present current approaches in a way that facilitates comparison of like with like in order to arrive at an informed perspective. The poster offers a framework that exposes certain key aspects of the parsing process when MWEs are involved. In this proposal we mainly discuss *orchestration*, which concerns the interplay between the activities of parsing and MWE recognition. Another aspect is the *MWE lexicon,* or more accurately, the manner in which information about MWEs is made available.

Introduction

This document attempts to present current research concerning the relation between the themes of hybrid parsing and the role of MWEs in the parsing process. There is imperfect agreement in the field about the nature of either of themes, let alone their intersection, so to document current approaches, we must initially set out a shared background against which the discussion can take place. To set the scene we first make general assumptions about the process of parsing itself. Most authors agree about the input/output characteristics of parsing - namely that it takes text in the form of a sentence consisting of words as input and outputs one or more competing representations of that sentence's structure as shown below.

The structures resulting from parsing vary enormously, ranging from shallow, surface-oriented chunks to deeper, more abstract representations that provide a stepping-stone to the underlying semantics. We postpone discussion of these structures in order to concentrate upon the issue of MWE recognition, since that is highly pertinent to the relationship between parsing and MWEs. In the schematic below, we take the simplest case whereby the input sentence contains an MWE that occupies contiguous positions i to j, marked in red. In more complex cases, the positions occupied by the MWE may not be contiguous, but we will ignore this for the moment.

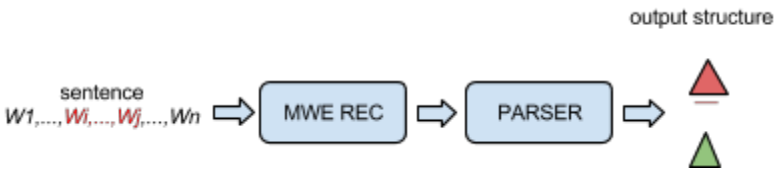


The small red line under the first representation signifies that the MWE has been recognised and incorporated into the output structure. However, in the other representation the MWE is not recognised. This is a standard case of ambiguity. Whichever reading is correct, the MWE has to be recognised in the first case thus raising two issues: (i) when the recognition takes place and (ii) how it takes place. We will refer to the first issue as *orchestration* with which this proposed poster is mainly concerned. The second issue concerns the storage and retrieval of information pertinent to the recognition and analysis of MWEs. In the above example, we need to know that the MWE is made up of two particular words and that the resulting MWE has a definite grammatical function. It seems inescapable that whatever the orchestration method, the existence of an MWE repository in some form is assumed and we will refer to this as the *lexicon* issue. We will address different approaches to genesis, representation and use of such lexicons more fully in a subsequent document.

Orchestration

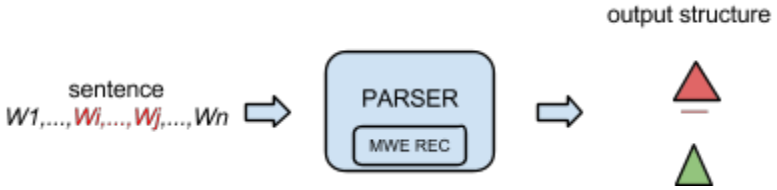
We consider three points at which recognition could take place: (i) before parsing (ii) during parsing and (iii) after parsing.

MWE recognition before parsing



When methods are used to partly annotate words and sequences of words before parsing, the search space of the parsing algorithm is reduced; hence the main advantage of MWE detection before parsing is that the parsing process becomes less complex. In that case, the parser takes as input a sequence (or lattice) of partially analyzed linguistic units. In most works using this strategy (e.g. Cafferkey et al. 2007, Korkontzelos and Manandhar 2010, Constant et al. 2012), a predicted MWE is merged into a single token (e.g. *by the way -> by_the_way*). Another strategy consists in reducing the predicted MWE into a subtree (formed of several nodes). For instance, in Schneider (2014), the MWE is reduced to a subtree whose root is the MWE head. The disadvantage of methods using prerecognition is that they are deterministic, so parsers cannot recover from their overgeneration. A sentence like *He recognises her by the way she walks* cannot be analysed correctly if '*by the way'* is pre-analyzed as an MWE adverb (*by_the_way*). To deal with this problem at least partially, Constant et al. (2013) use a MWE recognizer that generates the *most probable* outputs in the form of a lattice of linguistic units to be used as input of the parser. The parser is then in charge of selecting the best lexical segmentation. From these considerations a crucial aspect of MWE pre-recognition is whether ambiguous results be handled.

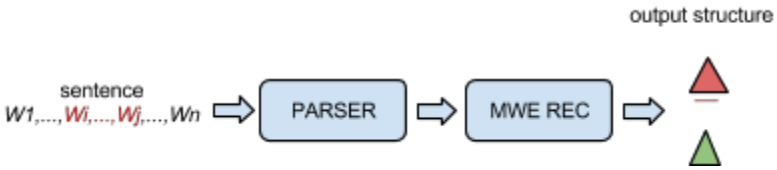
MWE recognition during parsing



If parsing is to profit from using MWE recognition from a pre-processing step, it has to be carried out non-deterministically, since several alternatives must be maintained and eventually resolved using a disambiguation model of some kind. For example, a statistical model might be used in collaboration with the parsing step. Various authors have employed such a joint approach using different parsing frameworks. In the dependency parsing framework, specific arcs annotate MWE components with either a shallow structure (Nivre and Nilsson 2004, Eryigit 2011, Seddah et al. 2013, Kong et al. 2014), or a deeper one. For instance, Vincze et al. (2013) and Candito and Constant (2014) incorporate syntactic functions in the MWE labels. In the constituent parsing framework, each MWE is annotated with a specific subtree (Arun and Keller 2005; Green et al. 2011, 2013). Also of note is the work of Finkel and Manning (2009) limited to Named Entity Recognition and constituent parsing.

MWEs can be divided into two semantic classes: those which are largely compositional (they are often called collocations, e.g. Named Entity Recognition), and those which are non-compositional (they are often called idioms, e.g. kick the bucket). For the former type, the classical parsing disambiguation approaches such as bi-lexical probabilities (e.g. Collins 1999) can be used, and semantic resources can improve results. For instance, Schneider (2012) reports e.g. that adding semantic expectations (including collocations) improves parsing performance on the (generally highly ambiguous) PP-attachment relations. Wehrli et al. (2010) also shows that collocation detection at parsing time improves parsing accuracy.

MWE recognition after parsing



The standard approach to detecting collocations is typically using an observation window of fixed length e.g. 3 or 5 words to the left and/or to the right and counts every word in this window as co-occurrence. It has been shown by several authors, for example Seretan (2011) or Bartsch and Evert (2014) that collocation extraction using syntactic relations leads to better results. Also parsing results themselves can still be improved by using re-ranking approaches (Charniak and Johnson, 2005). For instance, Constant et al. (2012) used a MWE-dedicated re-ranker on top of a parser generating the n-best parses (including MWE identification).

MWE lexicons

MWE prerecognition can be performed using several approaches: finite-state lexicon-based preprocessor, supervised labelers, unsupervised recognizers, or a combination of all. There are two ways of exploiting an MWE lexicon in a MWE-aware parsing system: (i) hard constraints: used to limit search space (Cafferkey et al. 2007, Korkontzelos and Manandar 2010), and (ii) soft constraints: used as source of features in discriminative models: during preprocessing stage (e.g. CRF-based MWE recognizer), at parsing time (SVM for dependency parsing, …), during postprocessing (e.g. reranker, …)

Discussion

Most of papers cited above present hybrid approaches. For instance, statistical parsers might be based based on grammatical formalisms: e.g. Context-free grammar (Arun and Keller 2005), Tree substitution grammars (Green et al. 2011, 2013), or based on a set of humanly designed "rules" (e.g. transitions in the Nivre (2014)'s proposal to integrate MWE recognition in a transition-based parser).

Statistical parsing systems may interact with different kinds of resources: e.g. semantic resources (Schneider 2014), MWE resources (e.g. Candito and Constant 2014).

Conclusion

This initial excursion into the classification of current work on hybrid parsing is clearly preliminary. We are aware that further input is required to complete the sections on the lexicon and on the nature of hybrid parsing. Nevertheless we are convinced the proposed poster will succeed in stimulating further group discussion that will enrich the forthcoming state-of-the-art report.

References

Arun A., Keller F. (2005), « Lexicalization in crosslinguistic probabilistic parsing : The case of French », Proceedings of the Annual Meeting of the Association For Computational Linguistics (ACL'05), p. 306-313

Bartsch, S. and Evert, S. (2014). Towards a Firthian notion of collocation. In A. Abel and L. Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, number 2/2014 in OPAL - Online publizierte Arbeiten zur Linguistik, pages 48-61. Institut für Deutsche Sprache, Mannheim.

Cafferkey C., Hogan D., van Genabith J. (2007), « Multi-word units in treebank-based probabilistic parsing and generation », Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP'07).

Candito M. and Constant M. (2014), Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. *52th Annual Meeting of the Association for Computational Linguistics (ACL'14)*

Charniak, E. and Johnson M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Constant M., Sigogne A.,Watrin P. (2012), « Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing », Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12), p. 204-212.

Constant M., Le Roux J. and Sigogne A. (2013). Combining Compound Recognition and PCFG-LA Parsing with Word Lattices and Conditional Random Fields. ACM Transactions on Speech and Language Processing. Special Issue on Multiword Expressions. Vol. 10(3). 24 pp.

Eryigit G., Ilbay T., Arkan Can O. (2011), « Multiword Expressions in Statistical Dependency Parsing», Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages (SPRML'11), p. 45-55.

Finkel J. R., Manning C. D. (2009), "Joint Parsing and Named Entity Recognition", Proceedings of the annual conference of NAACL- Human Language Technologies (NAACL-HLT'09), p. 326-334.

Green S., de Marneffe M.-C., Bauer J., Manning C. D. (2011), «Multiword Expression Identification with Tree Substitution Grammars : A Parsing tour de force with French », Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11), p. 725-735.

Spence G., de Marneffe M.-C., and Manning C. D. (2013). Parsing models for identifying multiword expressions. Computational Linguistics, 39(1):195–227.

Kong L. Schneider N., Swayamdipta S., Bhatia A., Dyer C., Smith N. (2014), A dependency Parser for Tweets, Proceedings of the conference on Empirical Methods for Natural Language Processing (EMNLP 2014)

Korkontzelos, I. et Manandhar, S. (2010). Can recognising multiword expressions improve shallow parsing ? In Proceedings of Human Language Technologies : The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'10), pages 636–644.

Nivre J., Nilsson J. (2004), «Multiword units in syntactic parsing », Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA).

Nivre, J. (2014). Transition-Based Parsing with Multiword Expressions (Athens WG3 poster)

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Springer, Dordrecht.

Seddah D., Tsarfaty R., Kübler S., Candito M., Choi J.D., Farkas R., Foster J., Goenaga I., Gojenola Galletebeitia K., Goldberg Y., Green S., Habash N., Kuhlmann M., Maier W., Marton Y., Nivre J., Przepiórkowski A., Roth R., Seeker W., Versley Y., Vincze V., Woliski M., Wróblewska A., Villemonte de la Clergerie, E. (2013), Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. Proceedings of the Fourth SPMRL Workshop, Seattle, USA

Schneider, G. (2012). Using semantic resources to improve a syntactic dependency parser. In Viktor Pekar Verginica Barbu Mititelu, Octavian Popescu, editor, *SEM-II workshop at LREC 2012*. Istanbul.

Schneider, G. (2014), Improving PP attachment in a hybrid dependency parser using semantic, distributional and lexical resources (Athens WG3 poster)

Veronika Vincze, Janos Zsibrita, and Istvan Nagy T. (2013). Dependency parsing for identifying hungarian light verb constructions. In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan.

Wehrli E., Seretan V., Nerima L. (2010), « Sentence analysis and collocation identification », Proceedings of the Workshop on Multiword Expression : From Theory to Applications (MWE'10), p. 28-36