Improving Natural Language Parsing Through Machine Learning and Lexical Resources

Conor Cafferkey

National Centre for Language Technology (NCLT)
School of Computing, Dublin City University

ccafferkey@computing.dcu.ie

Objective

Improve the accuracy of natural language parsing by employing machine learning (ML) techniques and exploiting existing lexical resources (machine-readable dictionaries, or MRDs).

Natural language parsing

Parsing is the process of deducing the syntactic structure of a string. It is the prerequisite for many natural language processing (NLP) tasks (Lease et al., 2006). It is used in applications such as Information Extraction (IE), Machine translation (MT) and Text Summarisation.

Types of parsing:

- Shallow, or 'skeletal', constituency parsing (Figure 1) (e.g. Collins, 1999)
- Deep parsing based on grammar formalisms such as Lexical-Functional Grammar (LFG)
- Dependency parsing (Figure 2) (Nivre, 2005)

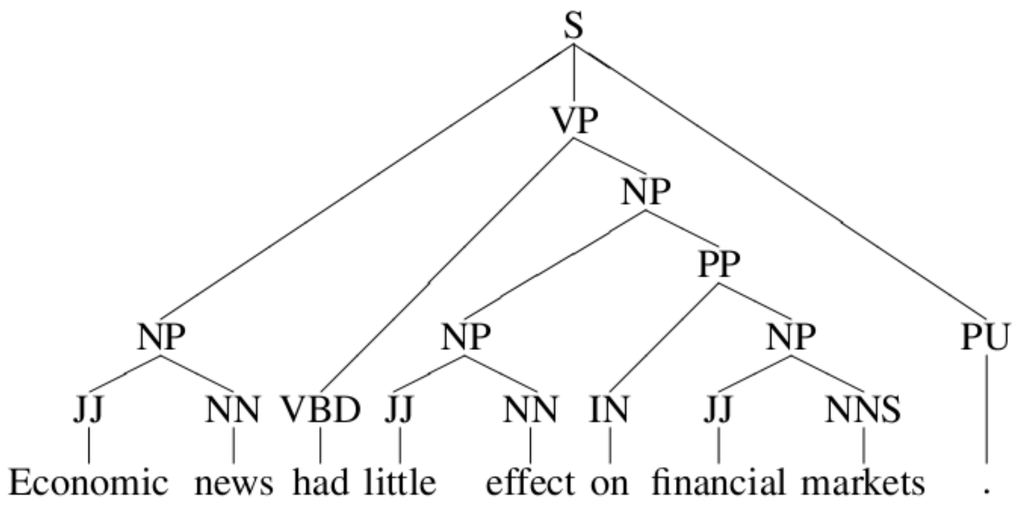

Figure 1: A 'shallow' constituency parse for the sentence *Economic news had little effect on financial markets* (from Nivre, 2005)

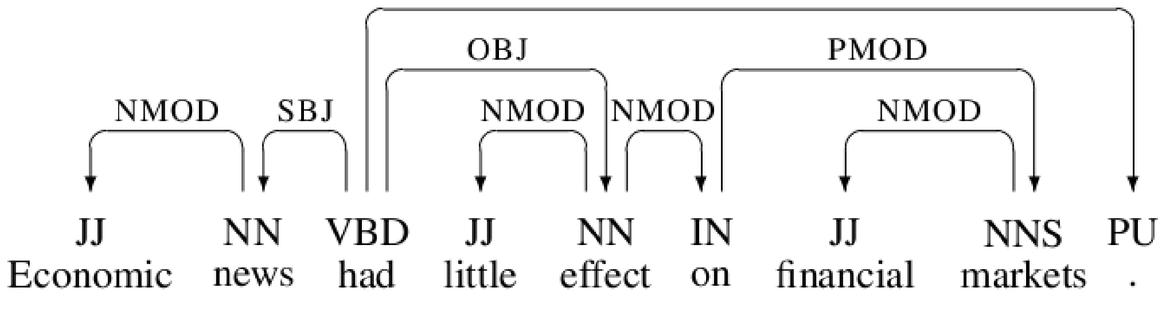

Figure 2: A dependency analysis for the same sentence (Nivre, 2005)

Natural language parsing is typically data-driven: trained on large-scale syntactically-annotated corpora such the Penn Treebank (Marcus et al., 1994).

ML techniques

ML techniques can be employed at several stages in parsing:

- Identifying partial parsing hypotheses prior to parsing to guide a probabilistic parser
- Re-ranking candidate analyses produced by the parser
- Inducing better grammars from the training examples

A hybrid approach combining (ML-based) dependency parsing and constituency parsing?

Lexical resources

There exist several wide-coverage lexical resources that encode selectional preferences such as verb subcategorisation frames and word classes (examples include WordNet, VerbNet, FrameNet and COMLEX.)

These resources could potentially be used to disambiguate difficult parsing decisions.

Work to date

- Identifying multi-word units (MWUs) such as multi-word named entities (NEs) and multi-word prepositional expressions prior to parsing
- Identifying non-overlapping syntactic chunks (constituting partial parsing hypotheses) to guide the parser

Publications

D. Hogan, **C. Cafferkey,** A. Cahill and J. van Genabith. 2007. Exploiting Multi-Word Units in History-Based Probabilistic Generation. To appear in *Proceedings of EMNLP-CoNLL 2007*. Prague, Czech Republic.

C. Cafferkey, D. Hogan and J. van Genabith. 2007. Multi-Word Units in Treebank-Based Probabilistic Parsing and Generation. To appear in *Proceedings of RANLP 2007*. Boverets, Bulgaria.

References

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

M. Lease, E. Charniak, M. Johnson and D. McClosky. 2006. A Look At Parsing and Its Applications. In *AAAI 2006.* Boston, Massachusetts.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.

Joakim Nivre. 2005. *Dependency grammar and dependency parsing*. Technical Report. Växjö University, Sweden.

Supervisors: Prof. Josef van Genabith and Dr. Deirdre Hogan

Supported by the Irish Research Council for Science, Engineering and Technology (IRCSET) Embark Initiative and Microsoft Research Ltd.