Journal of Machine Learning Research 2 (2002) 551-558 Submitted 3/02; Published 3/02

Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing

James Hammerton J.HAMMERTON@LET.RUG.NL
Alfa-Informatica
University of Groningen
The Netherlands

Miles Osborne OSBORNE@COGSCI.ED.AC.UK
Division of Informatics
University of Edinburgh
Scotland

Susan Armstrong SUSAN.ARMSTRONG@ISSCO.UNIGE.CH
ISSCO/ETI
University of Geneva
Switzerland

Walter Daelemans WALTER.DAELEMANS@UIA.UA.AC.BE
Center for Dutch Language and Speech
University of Antwerp
Belgium

Abstract

This article introduces the problem of partial or shallow parsing (assigning partial syntactic structure to sentences) and explains why it is an important natural language processing (NLP) task. The complexity of the task makes Machine Learning an attractive option in comparison to the handcrafting of rules. On the other hand, because of the same task complexity, shallow parsing makes an excellent benchmark problem for evaluating machine learning algorithms. We sketch the origins of shallow parsing as a specific task for machine learning of language, and introduce the articles accepted for this special issue, a representative sample of current research in this area. Finally, future directions for machine learning of shallow parsing are suggested.

1. Introduction

In full parsing, a grammar and search strategy are used to assign a complete syntactic structure to sentences. The main problem here is to select the most plausible syntactic analysis given the often thousands of possible analyses a typical parser with a sophisticated grammar may return. Stochastic approaches can be used to order the analyses according to their probability or to generate the most probable parse(s) only. See Jurafsky and Martin (2000) for an introduction to traditional and stochastic approaches to parsing.

However, not all natural language processing (NLP) applications require a complete syntactic analysis. A full parse often provides more information than needed and sometimes less. E.g., in Information Retrieval, it may be enough to find simple NPs (Noun Phrases) and VPs (Verb Phrases). In Information Extraction, Summary Generation, and Question Answering, we are interested especially in information about specific syntactico-semantic relations such as agent, object, location, time, etc (basically, who did what to whom, when, where and why), rather than elaborate configurational syntactic analyses.

Partial or shallow parsing — the task of recovering only a limited amount of syntactic information from natural language sentences — has proved to be a useful technology for written and spoken language domains. For example, within the Verbmobil project, shallow parsers were used to add robustness to a large speech-to-speech translation system (Wahlster, 2000). Shallow parsers are also typically used to reduce the search space for full-blown, 'deep' parsers (Collins, 1996). Yet another application of shallow parsing is question-answering on the World Wide Web, where there is a need to efficiently process large quantities of (potentially) ill-formed documents (Buchholz and Daelemans, 2001, Srihari and Li, 1999). And more generally all text mining applications, e.g. in biology (Sekimizu et al., 1998).

Abney (1991) is credited with being the first to argue for the relevance of shallow parsing, both from the point of view of psycholinguistic evidence and from the point of view of practical applications. His own approach used hand-crafted cascaded Finite State Transducers to get at a shallow parse.

Typical modules within a shallow parser architecture include the following:

1. Part-of-Speech Tagging. Given a word and its context, decide what the correct morphosyntactic class of that word is (noun, verb, etc.). POS tagging is a well-understood problem in NLP (van Halteren, 1999), to which machine learning approaches are routinely applied.

2. Chunking. Given the words and their morphosyntactic class, decide which words can be grouped as chunks (noun phrases, verb phrases, complete clauses, etc.)

3. Relation Finding. Given the chunks in a sentence, decide which relations they have with the main verb (subject, object, location, etc.)

Because shallow parsers have to deal with natural languages in their entirety, they are large, and frequently contain thousands of rules (or rule analogues). For example, a rule might state that determiners (words such as *the*) are good predictors of noun phrases. These rule sets also tend to be largely 'soft', in that exceptions abound. Continuing with our example, in the phrase:

> ... fatalities on non-interstate roads were about the same

the word *the* is instead within the adjectival phrase *were about the same*. This example was taken from the Parsed Wall Street Journal (Marcus et al., 1993).

Building shallow parsers is therefore a labour-intensive task. Unsurprisingly, shallow parsers are usually automatically built, using techniques originating within the machine learning (or statistical) community.

The work by Ramshaw and Marcus (1995) proved to be an important inspiration source for this work. By formulating the task of NP-chunking as a tagging task, a large number of machine learning techniques suddenly became available to solve the problem. In this approach, each word is associated with one of three tags: I (for a word inside an NP), O (for outside of an NP), and B (for between the end of one and the start of another NP). The classification task can easily be extended to other types of chunks and with some effort even to finding relations (Buchholz et al., 1999). For an extension of a HMM approach from tagging to chunking, see Skut and Brants (1998).

Readers are encouraged to visit the *Computational Natural Language Learning* (CoNLL) shared task websites:[1]

http://lcg-www.uia.ac.be/conll2000/chunking/

and:

http://lcg-www.uia.ac.be/conll2001/clauses/

for background reading, datasets and results of more than 20 shallow parsing systems.

Applying learning techniques is however not necessarily straightforward:

- The amount of data to be processed will push batch systems to the limit. This means that learners will need to scale.
- Labelled training material is frequently noisy and only exists in relatively small quantities. Here, 'small' is with respect to a language as a whole. Any learner must therefore deal with overfitting.
- Real-world sentences tend to be long. Learners which do not operate in (near) linear time are simply unfit for the task.

Shallow parsing, like much of natural language processing, is therefore a challenging domain for machine learning research.

Note that shallow parsing does not refer to a single technique. Instead, it is better to consider it to refer to a family of related methods, all of which attempt to recover some syntactic information, at the possible expense of ignoring all other such information.

2. Overview of Papers

Here we briefly summarise the papers in this issue.

2.1 Memory Based Shallow Parsing

Tjong Kim Sang (2002) considered the issues involved with applying memory-based learning (MBL) to shallow parsing. MBL consistently performs well for a variety of shallow parsing tasks, often yielding (near) best results (Daelemans et al., 1999, Buchholz et al., 1999). From this, one might conclude that MBL was a promising learning technique for pushing

1. CoNLL is the yearly conference of SIGNLL, the Special Interest Group of the Association for Computational Linguistics on Machine Learning of Language; `http://www.aclweb.org/signll`.

shallow parsing to *full* parsing. For full parsing, MBL fared less well, however, and the results were not as good as for the other parsers that were compared. This does not mean that MBL is fundamentally unsuited for full-blown parsing. Instead, it suggests that the task needs to be encoded in some other manner.

In his paper, a weakness of MBL — that it can have difficulty handling large numbers of features — was identified. A feature selection method, namely bidirectional hill climbing (Caruana and Freitag, 1994), was found to yield insignificant gains in performance for NP parsing. However, it did produce a significant improvement for clause identification.

Tjong Kim Sang also showed how ensemble learning techniques such as (weighted) majority voting and stacking could improve upon performance. All system combination methods improved on the results of the individual MBL classifiers, and the best performer was to employ MBL itself as a stacked classifier.

2.2 Shallow Parsing using Specialized HMM

Molina and Pla (2002) presented a shallow parser based on Hidden Markov Models (HMMs). HMMS are routinely used in speech recognition and part-of-speech tagging (POS tagging). Here, the HMM was used to find the most probable sequence of output shallow parsing labels for the current sequence of inputs. Unlike with the previous MBL approach to shallow parsing (which is classification-based), this approach used a generation approach. Their generative model enabled information about the whole sentence to be taken into account when determining the output shallow parsing label for each word, since it is the probability of the whole sequence of output tags occurring given the current input that is maximised (and not just the probability of individual decisions). The authors' HMMs are applied to a variety of shallow parsing tasks.

Various ways of encoding the task were shown to produce different results. Clearly, this suggests that feature specification is an important issue. Interestingly enough, the authors, whilst not using ensemble learning methods, produced results comparable with systems which did use such techniques. Here, the obvious comparison is with the MBL paper mentioned in section 2.1. An interesting possibility here is that their generative model (which allows previous decisions to directly influence future decisions) emulates the ability of ensemble learners to correct for classifiers which do not take previous decisions into account.

2.3 Text Chunking Based on a Generalization of Winnow

Zhang et al. (2002) presented a generalised version of the Winnow algorithm. They observed that the original Winnow algorithm is only guaranteed to converge on linearly separable data. So, given the possibility that features for shallow parsing are not linearly separable, the authors modified Winnow such that it would converge, even for non-linearly separable features. They also showed that both versions of Winnow were robust to irrelevant features.

The authors used a very large set of features, including those derived from sources other than the training set. Winnow was found to be a strong performer for this task, giving the best results reported for a non-ensemble classifier in the CoNLL 2000 shared task. Clearly, the ability to exploit very large numbers of (potentially irrelevant) features is a crucial component of a successful shallow parsing system.

2.4 Shallow Parsing With PoS Taggers and Linguistic Knowledge

Megyesi (2002) retrained three POS taggers for shallow parsing. Unlike the other papers, she dealt with shallow parsing for Swedish, and not English.

Experimental results showed that, again, when using POS taggers as the basis of shallow parsers, careful consideration needs to be given to how the task is to be encoded (choice of features). Unlike other studies, the author found that ignoring lexical information improved performance for all her systems. It is unclear whether this is due to linguistic differences between English and Swedish, or else due to the fact that some of her POS taggers were built with English in mind.

The shallow parsers were then trained on varying amounts of training data for each task. Unsurprisingly performance improved with the amount of training data in each case. However, no shallow parser yielded uniformly superior results to any other shallow parser.

2.5 Learning Rules and their exceptions

Dejean (2002) presented a top-down rule induction system, called ALLiS, for learning linguistic structures. The initial system is enhanced with additional mechanisms to deal with noisy data. The author identifies two types of difficulties – significant noise in the data and the presence of linguistically motivated exceptions. Since linguistically motivated exceptions occur, they cannot be treated as noise. To address these problems, a refinement algorithm is introduced to learn exceptions for each rule that is learned. The second improvement introduces linguistically motivated prior knowledge to improve the efficiency and accuracy of the system.

The experimental results clearly demonstrate significant improvement with the introduction of the two mechanisms. The refinement mechanism is based on the assumption that there is some regularity to the errors in the data and thus, by systematically searching for exceptions, the rule induction system is improved. With the use of prior knowledge, the context of only one element need be taken into account and the search space is reduced resulting in a significant reduction in learning time. In comparison to (Brill, 1994), a well-known transformation based learning system (TBL), ALLiS needs fewer rules and overcomes a number of classification errors produced by TBL.

The incorporation of linguistically motivated prior knowledge in a learning-based system is an interesting addition, and as pointed out in the paper, the question arises whether such background information would be useful in other systems. In any case, it is clear that additional mechanisms are necessary to deal with the noise and exceptions present in natural language data for tasks such as shallow parsing.

2.6 Shallow Parsing using Noisy and Non-Stationary Training Material

Osborne (2002) considered an issue that has gone largely unaddressed in the shallow parsing literature, namely what happens when the training set is either noisy, or else drawn from a different distribution to the testing material.

This paper took a range of shallow parsers (including both single model parsers and ensemble parsers) and trained them using various types of artificially noisy material. In a second set of experiments, the issue of whether naturally occurring disfluencies have more impact on performance than a change in the distribution of the training material was investigated. It was found that the changes in the distribution are more important.

The author drew various conclusions from this work. Shallow parsers are robust and only large quantities of noise will significantly impair performance. Should one wish to improve performance then simple parser specific extensions can help. No single technique worked best with all types of noise with different kinds of noise favouring different parsers. Regarding the results on changes in the distribution of training data, the clear lesson is that if one wishes to improve the performance of shallow parsers on a particular task, it is better to annotate more examples from the target distribution than to use additional training material from other distributions.

One surprise in this paper is that the parsers employing system combination, although generally the best performers in the literature, were not always the best at dealing with noise. Clearly, ensemble learning is not always a sure-fire strategy.

3. Conclusions

In summary, a few points can be made:

- Feature selection, as in machine learning in general, is an important consideration for machine learning of shallow parsers. Some learning approaches only work well when the features have been carefully selected and weighted, whilst others can cope with large numbers of irrelevant features. The Winnow and MBL papers both clearly illustrated these considerations.

- A recent trend in the literature is for performance to be potentially improved by training several classifiers on the task and combining their results to produce a final result. This can be done in various ways such as using various (weighted) voting methods and using stacked classifiers. This however is not guaranteed to produce the best results as Osborne's paper above illustrates.

- The majority of the systems are probabilistic, with the obvious exception of MBL. Few shallow parsers reported in the literature are, for example based upon Inductive Logic Programming or neural networks. It seems that the reason for this is the need for scalability.

- All parsers assumed labelled input. Clearly this limits performance, as only a small amount of labelled training material exists. Zhang et al. (2002) did use other knowledge sources, in addition to the training set.

- Shallow parsers are noise-tolerant, and only massive quantities of noise will significantly undermine performance.

- Not all shallow parsers used generative models (as might be expected from the nature of the task). Discriminative models (those which attempt to maximise the difference between alternative labels, but not necessarily model the distribution of annotated sentences) are also employed. However, the exact link between these two classes of models has yet to be demonstrated.

Research in shallow parsing is clearly ongoing. We hope that more machine learning researchers will take-up the gauntlet and include shallow parsing as an additional, real-world domain with which to evaluate machine learning systems.

Acknowledgements

The editors wish to thank the following reviewers for their valuable help in producing this special issue:

> Richard K. Belew, Thorsten Brants, Eric Brill, Mary Califf, Claire Cardie, Rafael Carrasco, Alexander Clark, Steve Clark, Daniel Gildea, Hans Van Halteren, Jamie Henderson, Colin De La Higuera, Yuval Krymolowski, Marshall Mayberry, Grace Ngai, Adwait Ratnaparkhi, Erik F. Tjong Kim Sang, Jimi Shanahan, Cindi Thompson, Chris Watkins, Ton Weijters and Maria Wolters.

All of the editors were involved, at some time or other, with the EU TMR project *Learning Computational Grammars*. The homepage for the project is at http://lcg-www.uia.ac.be/. We wish to thank John Nerbonne for leading the project, Erik F. Tjong Kim Sang for maintaining the excellent shallow parsing website and finally the editorial staff of the JMLR for supporting this special issue.

References

S. Abney. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, Dordrecht, 1991.

E. Brill. Some advances in rule-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Seattle, Washington*, 1994.

Sabine Buchholz and Walter Daelemans. Complex Answers: A Case Study using a WWW Question Answering System. *Natural Language Engineering*, 2001.

Sabine Buchholz, Jorn Veenstra, and Walter Daelemans. Cascaded grammatical relation assignment. In Pascale Fung and Joe Zhou, editors, *Proceedings of EMNLP/VLC-99*, pages 239–246. ACL, 1999.

R. Caruana and D. Freitag. Greedy attribute selection. In *Proceeedings of the Eleventh International Conference on Machine Learning*, pages 28–36, New Brunswick, NJ, USA, 1994. Morgan Kaufman.

Michael John Collins. A new statistical parser based on bigram lexical dependencies. In 34^{th} *Annual Meeting of the Association for Computational Linguistics*. University of California, Santa Cruz, California, USA, June 1996.

W. Daelemans, S. Buchholz, and J. Veenstra. Memory-based shallow parsing. In *Proceedings of CoNLL*, Bergen, Norway, 1999.

H. Dejean. Learning rules and their exceptions. *Journal of Machine Learning Research*, 2002.

D. Jurafsky and J. Martin. *Speech and Language Processing.* Prentice-Hall, 2000.

M. Marcus, S. Santorini, and M. Marcinkiewicz. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL `citeseer.nj.nec.com/marcus93building.html`.

B. Megyesi. Shallow parsing with pos taggers and linguistic knowledge. *Journal of Machine Learning Research*, 2002.

A. Molina and F. Pla. Shallow parsing using specialized hmm. *Journal of Machine Learning Research*, 2002.

M. Osborne. Shallow parsing using noisy and non-stationary training material. *Journal of Machine Learning Research*, 2002.

L.A. Ramshaw and M.P. Marcus. Text chunking using transformation-based learning. In *Workshop on Very Large Corpora*, pages 82–94, 1995.

T. Sekimizu, H. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts, 1998.

W. Skut and T. Brants. Chunk tagger: statistical recognition of noun phrases. In *ESSLLI-1998 Workshop on Automated Acquisition of Syntax and Parsing*, 1998.

R. Srihari and W. Li. Information extraction supported question answering. In *Proceedings of TREC 8*, 1999.

E.F. Tjong Kim Sang. Memory-based shallow parsing. *Journal of Machine Learning Research*, 2002.

H. van Halteren. *Syntactic Wordclass Tagging.* Kluwer Academic Publishers, 1999.

Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-toSpeech Translation.* Springer, 2000.

T. Zhang, F. Damereau, and D. Johnson. Text chunking base on a generalization of winnow. *Journal of Machine Learning Research*, 2002.