Reliability in Example-Based Parsing

Oliver Streiter

Academia Sinica, Institute of Information Science, Nankang, Taipei, Taiwan 115
oliver@hp.iis.sinica.edu.tw http://rockey.iis.sinica.edu.tw

Abstract

In this paper we introduce an example-based parser for Chinese. One strong point of the parsers is its high reliability. We propose a formal definition for reliability and derive from it K as a metric for the evaluation of parsers. In a row of experiments we try to identify some factors which support the reliability of the parser. It is suggested that these factors are independent of the parsing approach and can be realized in TAGs.

1. Introduction

Example-based parsers adhere to the *lazy learning algorithm* while converting tree-bank entries into a parser. So-called *treebank grammars*, (Bod, 1992; Charniak, 1996) are *eager learners*, i.e. they abstract knowledge structures or statistical information from the treebank and reason on the basis of these abstractions. *Explanation-based parsing* is a different eager learning approach aiming at the extraction of specialized grammars out of a general-purpose grammars on the bases of parsing examples (Rayner & Christer, 1994; Srivinas & Joshi, 1995).

Lazy learners keep all training data (e.g. all trees in the treebank) available in their original form. They may operate on similar abstractions as *eager learners* do, e.g. parse from partial trees with category labels, but dispose in addition of the original encoding which can be referred to if generalizations become ambiguous (Daelemans *et al.*, 1999). The learning set is not filtered or modified and contains among regular phenomena redundancies, syntactic and semantic exceptions, phraseologies including lexical functions (Mel'čuk, 1974), pronouns with their antecedents, markers of text-coherence (e.g. *fire, cigarette, match*), and pieces of common sense knowledge (*he sees the sparrow with the spyglass*), all pieces of information which are necessary, or at least helpful for high-quality parsing (Doi & Maraki, 1992; Bod, 1999).

All words and categories are of equal importance to the parser unless special weights are assigned to them. It might be argued that this equal distribution of weights is not sense-less and that, for example, the linguistic notion of head as pivot should and can be dispensed with. Giving preference to specific matches (e.g. verbs) might produce a bias which endangers the reliability, i.e. a good match is not chosen, just because another match contains more verbs. Linguistic support may come from observations in verb-last languages where speakers are contradicted/approved before the final main verb has been pronounced. The list of actants, circumstances, lexical functions as magnifiers etc are often sufficient in order to identify the verb or its syntactic or semantic type.

2. An Example-Based Parser

An example-based parser is currently developed at the Academia Sinica of Taiwan (Streiter, 1999; Streiter & Hsueh, 2000), based on a Chinese treebank of about 30.000 trees (Chen *et al.*, 1999). The annotation scheme comprises 200 lexical labels, 45 phrasal categories and 46 semantic roles. The parser retrieves trees from a treebank via a fuzzy match of the sentence to be

parsed and the terminals of the all trees in the treebank. The 20 best matching trees are further processed and aligned with the sentence in case the tree is smaller than the sentence. The best aligned tree is selected. Mainly through re-parsing awkward subtrees, badly matched trees are corrected and unmatched words are inserted. The parser is fast and by means of the fuzzy match extremely robust. The complexity of other parsing approaches is avoided, as parsing consists mainly of retrieving large chunks from a databank. The coverage, as evaluated in (Streiter & Chen, 2000) is not yet fully satisfying. Unchallenged, however, is the reliability of this parser.

3. What Reliability is about

Reliability is an important evaluation criterion for NLP which until now has failed to obtain a formal definition as well the attention it merits. The standard evaluation tests a parser on unlearned corpora, determining its coverage in terms of recall and precision. The pendant of the coverage is the *reliability*, which we define as a system property, i.e. as *performance on trained corpora*. Reliability is thus close the notion of *tunability*. However, the impact of reliability is more fare-reaching: A system which has a high reliability can always enlarge its coverage by learning new items. A system with low reliability cannot improve its coverage by learning new items: the system is quickly over-trained.

We define coverage ($\mathcal{C}$) and reliability ($\mathcal{R}$) as meta-scores which elaborate the values of recall and precision. As $\mathcal{R}$ is neither compatible with low precision (false alarm) nor with low recall (a silent system), we define $\mathcal{R}$ and $\mathcal{C}$ as f-score with learned respectively unlearned test corpora. With this definitions we formulate the *hypothesis of converging $\mathcal{C}$ and $\mathcal{R}$*: **1)** $\mathcal{R}$ is always higher than $\mathcal{C}$. **2)** $\mathcal{R}$ decreases with more training data (due to ambiguities which arise). **3)** $\mathcal{C}$ approaches $\mathcal{R}$ with more training data (more items are known or similar to known items). **4)** Before $\mathcal{C}$ and $\mathcal{R}$ converge $\mathcal{C}$ may decreases under the influence of decreasing $\mathcal{R}$.

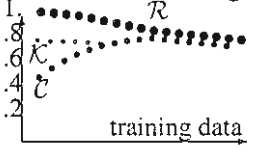


Figure 1: The hypothesis of converging $\mathcal{C}$ and $\mathcal{R}$ for hypothetical data. $\mathcal{K}$ is the estimated maximal coverage.

While most experimental data available support an asymptotic rise of $\mathcal{C}$, little is known about $\mathcal{R}$. Given the above (hypothetic) distribution, the maximal coverage a system can achieve as well as its current position are important data. We propose to estimated the *maximal coverage* $\mathcal{K}$ as $\mathcal{C} + \frac{(\mathcal{R}^2-\sqrt{\mathcal{C}})\cdot\sqrt{\mathcal{C}}}{(1-\mathcal{R}^2)+\sqrt{\mathcal{C}}}$. With $\mathcal{K} > \mathcal{C}$ further investment in more teaching is profitable, otherwise system properties have to be changed in order to enforce $\mathcal{R}$ and with it future grow.

4. Factors determining Reliability

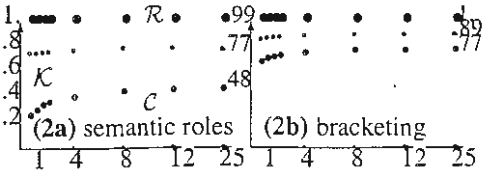
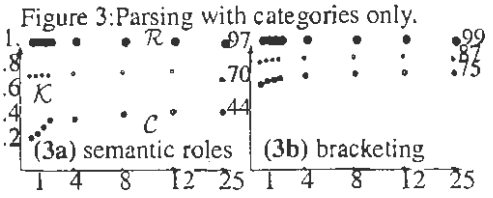


Figure 2: **Reference Data**. $\mathcal{C}$ and $\mathcal{R}$ for 500 to 25.000 training sentences, evaluating the recognition of semantic relations (agent, theme, goal, experiencer, time etc) between head and dependents (2a) and the bracketing (2b) for the example-based parser on a test-corpus of 727 sentences.

Experiment 1 In order to establish the effect of the string and lexeme encoding in addition to the category encoding we removed the string and lexeme encoding as done in all eager learners.

Figure 3:Parsing with categories only.

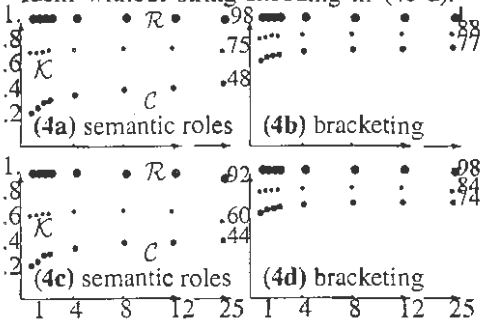


F.(3) shows a loss is $\mathcal{R}$, $\mathcal{C}$ and $\mathcal{K}$ compared to F.(2). We assume that the drop in $\mathcal{C}$ has been produced by the drop in $\mathcal{R}$, as unknown items are treated only in reference to learned items. It is the ambiguity in the learned items (the inverse of $\mathcal{R}$) which causes the drop of $\mathcal{C}$.

Experiment 2 In order to establish the effect of the context sensitivity we not only re-parsed awkward subtrees (see our description of the parser above), but re-parsed (artificially) all subtrees, thus breaking the links between sisters.

We observe a small loss of $\mathcal{R}$ compared to F.(2). If we test a context-free version with category encoding only (4c-d) and thus simulate standard parsing approaches, we observe an additional drop of $\mathcal{R}$ compared to F.(3). Thus context sensitivity is important for $\mathcal{R}$ but to a smaller extend than the encoding of lexemes and strings. Without string encoding the context-free grammar loses heavily in its $\mathcal{R}$. The drop of $\mathcal{K}$ shows that the loss cannot be compensated for by more training data.

Figure 4: Context-free parsing in (4a-b). Idem without string-encoding in (4c-d).



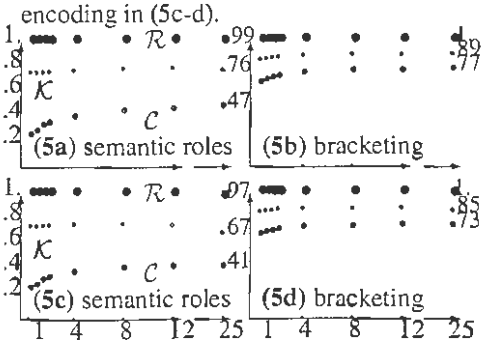
Experiment 3 To test the equal distribution of weights, we assigned 0.5 points for a matching verb, in addition to the 1 point for every match, assuming that in most cases the verb functions as head and a matching head is more important than a matching non-head.

F.(5) shows a small loss of $\mathcal{C}$ for semantic roles compared to F.(2). With category encoding only, we observe a drop of $\mathcal{C}$ for the bracketing compared to (3). The drop in the bracketing supports our claim that the bias is towards matching deeper branching structures by preference. This bias is unlikely to be produced by the specific additional score 0.5 we assigned:

score	+0	+0.25	+0.5	+1
K	.8660	.8585	.8548	.8393
Figure	3b		5d	

Figure 6: $\mathcal{K}$ for bracketing with additional scores to verbs when parsing without string-encoding (25.000 training sentences).

Figure 5:Additional scores to verb matches in (5a-b). Idem without string-encoding in (5c-d).



5. Summary

We have introduced, although shortly, an example-based parser. A formal grammar which bears most resemblance to this approach is TAG. Both approaches are based on collections of trees, atomic trees for TAGs and all trees and subtrees for example-based grammars. Parsing starts similarly by extracting trees via the indices formed by words. A distinguishing property of example-based grammars is that a tree preserves all terminal nodes per tree. The influence

of this strategy could not be tested, as this would require to leave the paradigm suggested. However, we could evaluate the effect of not necessarily distinguishing features, (i.e. the string-lemma encoding, the high degree of context-sensitivity and the non-preference of heads.)
The experiments have been preceded by a discussion of the notions of $\mathcal{R}$ and $\mathcal{C}$, for which a formal definition has been proposed. $\mathcal{K}$ has been proposed as evaluation measure which is less dependent on the size of the training corpus than $\mathcal{R}$ and $\mathcal{C}$ are.
In the experiments we could show that the string-lemma encoding is of utmost importance for $\mathcal{R}$ and $\mathcal{C}$, even though a very rich set of categories is employed. When the string encoding is renounced to, the grammar becomes more dependent on other features, such as a high degree of context-sensitivity and the correct assignment of weights.
The dominant role the head plays in formal grammars has been questioned as it has no priority in parsing relevant dimensions such as world knowledge, text coherence and idiomaticity.
Throughout 14 meaningful comparisons of test settings we observe 12 cases in which $\mathcal{R}$ and $\mathcal{C}$ decrease both. In two instance $\mathcal{C}$ improved with $\mathcal{R}$ remaining equal or decreasing, thus supporting our claim of a causal relation between declining $\mathcal{R}$ and declining $\mathcal{C}$.

6. Conclusion

Example-based grammars base their $\mathcal{R}$ mainly on the string-encoding. We hypothesize that TAGs with multiple terminals and a string-lemma encoding, if still be called TAG, could handle NLP task more reliable. In order to achieve this, automatic learning experiments should apply, unlike past experiments (Srivinas & Joshi, 1995; Xia, 1999), lazy learning approaches.

References

BOD R. (1992). Data oriented parsing (DOP). In *COLING*.

BOD R. (1999). Extracting stochastic grammars from treebanks. In *Journées ATALA sur les Corpus annotés pour la syntaxe*: Talana, Paris VII.

CHARNIAK E. (1996). Tree-bank grammars. In *13th National Conference on Artificial Intelligence*.

CHEN K.-J. ET AL. (1999). The CKIP Chinese Treebank. In *Journées ATALA sur les Corpus annotés pour la syntaxe*: Talana, Paris VII.

DAELEMANS W., BUCHHOLZ S. & VEENSTRA J. (1999). Memory-based shallow parsing. In *Proceedings of CoNLL-99*, Bergen, Norway. //ilk.kub.nl/papers.html.

DOI S. & MARAKI K. (1992). Translation ambiguity resolution based on text corpora of source and target language. In *COLING'92*.

MEL'ČUK I. A. (1974). *Opyt teorii lingvističeskix modelej Smysl⇔Tekst. Semantika, sintaksis*. Moskva

RAYNER M. & CHRISTER S. (1994). Corpus-based grammar specification for fast analysis. In A. ET AL., Ed., *Spoken Language Translator: First Year Report*, SRI Technical Report CRC-043.

SRIVINAS B. & JOSHI A. K. (1995). Some novel applications of explanation-based learning to parsing lexicalized tree-adjoining grammars. cmp-lg archive 9505023.

STREITER O. (1999). Parsing Chinese with randomly generalized examples. In *NLPRS'99 Workshop on Multi-lingual Information Processing and Asian Language Processing*, Beijing.

STREITER O. & CHEN K.-J. (2000). Experiments in example-based parsing. In *Dialogue 2000, International Seminar in Computational Linguistics and Applications*, Tarusa, Russia.

STREITER O. & HSUEH P.-Y. (2000). A case-study on example-based parsing. In *International Conference on Chinese Language Computing 2000*, Chicago.

XIA F. (1999). Extracting TAGs from bracketed corpora. In *Proceedings NLPRS'99*. Beijing.