Joint Parsing and Semantic Role Labeling

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Joint Parsing and Semantic Role Labeling

Charles Sutton and Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA
{casutton,mccallum}@cs.umass.edu

Abstract
A striking feature of human syntactic processing is that it is context-dependent, that is, it seems to take into account semantic information from the discourse context and world knowledge. In this paper, we attempt to use this insight to bridge the gap between SRL results from gold parses and from automatically-generated parses. To do this, we jointly perform parsing and semantic role labeling, using a probabilistic SRL system to rerank the results of a probabilistic parser. Our current results are negative, because a locally-trained SRL model can return inaccurate probability estimates.

1 Introduction
Although much effort has gone into developing statistical parsing models and they have improved steadily over the years, in many applications that use parse trees errors made by the parser are a major source of errors in the final output. A promising approach to this problem is to perform both parsing and the higher-level task in a single, joint probabilistic model. This not only allows uncertainty about the parser output to be carried upward, such as through an $k$-best list, but also allows information from higher-level processing to improve parsing. For example, Miller et al. (2000) showed that performing parsing and information extraction in a joint model improves performance on both tasks. In particular, one suspects that attachment decisions, which are both notoriously hard and extremely important for semantic analysis, could benefit greatly from input from higher-level semantic analysis.

The recent interest in semantic role labeling provides an opportunity to explore how higher-level semantic information can inform syntactic parsing. In previous work, it has been shown that SRL systems that use full parse information perform better than those that use shallow parse information, but that machine-generated parses still perform much worse than human-corrected gold parses.

The goal of this investigation is to narrow the gap between SRL results from gold parses and from automatic parses. We aim to do this by jointly performing parsing and semantic role labeling in a single probabilistic model. In both parsing and SRL, state-of-the-art systems are probabilistic; therefore, their predictions can be combined in a principled way by multiplying probabilities. In this paper, we rerank the $k$-best parse trees from a probabilistic parser using an SRL system. We compare two reranking approaches, one that linearly weights the log probabilities, and the other that learns a reranker over parse trees and SRL frames in the manner of Collins (2000).

Currently, neither method performs better than simply selecting the top predicted parse tree. We discuss some of the reasons for this; one reason being that the ranking over parse trees induced by the semantic role labeling score is unreliable, because the model is trained locally.

2 Base SRL System
Our approach to joint parsing and SRL begins with a base SRL system, which uses a standard architecture from the literature. Our base SRL system is a cascade of maximum-entropy classifiers which select the semantic argument label for each constituent of a full parse tree. As in other systems, we use three stages: pruning, identification, and classification. First, in pruning, we use a deterministic pre-processing procedure introduced by Xue and Palmer (2004) to prune many constituents which are almost certainly not arguments. Second, in identification, a binary MaxEnt classifier is used to prune remaining constituents which are predicted to be null with
Table 1: Features used in base identification classifier.

| Base features [GJ02] | Path to predicate | Constituent type | Head word | Position | Predicate | Head POS [SHWA03] | All conjunctions of above |

Table 2: Features used in baseline labeling classifier.

<table>
<thead>
<tr>
<th>Parse Trees Used</th>
<th>SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>77.1</td>
</tr>
<tr>
<td>1-best</td>
<td>63.9</td>
</tr>
<tr>
<td>Reranked by gold parse F1</td>
<td>68.1</td>
</tr>
<tr>
<td>Reranked by gold frame F1</td>
<td>74.2</td>
</tr>
<tr>
<td>Simple SRL combination ($\alpha = 0.5$)</td>
<td>56.9</td>
</tr>
<tr>
<td>Chosen using trained reranker</td>
<td>63.6</td>
</tr>
</tbody>
</table>

Table 3: Comparison of Overall SRL F1 on development set by the type of parse trees used.

In this section we present results on several reranking methods for joint parsing and semantic role la-

3 Reranking Parse Trees Using SRL Information

Here we give the general framework for the reranking methods that we present in the next section. We write a joint probability model over semantic frames $F$ and parse trees $t$ given a sentence $x$ as

$$p(F, t|x) = p(F|t, x)p(t|x),$$

where $p(t|x)$ is given by a standard probabilistic parsing model, and $p(F|t, x)$ is given by the baseline SRL model described previously.

In this paper, we choose $(F^*, t^*)$ to approximately maximize the probability $p(F, t|x)$ using a reranking approach. To do the reranking, we generate a list of $k$-best parse trees for a sentence, and for each predicted tree, we predict the best frame using the base SRL model. This results in a list $\{(F_i, t_i)\}$ of parse tree / SRL frame pairs, from which the reranker chooses. Thus, our different reranking methods vary only in which parse tree is selected; given a parse tree, the frame is always chosen using the best prediction from the base model.

The $k$-best list of parses is generated using Dan Bikel’s (2004) implementation of Michael Collins’ parsing model. The parser is trained on sections 2–21 of the WSJ Treebank, which does not overlap with the development or test sets. The $k$-best list is generated in Bikel’s implementation by essentially turning off dynamic programming and doing very aggressive beam search. We gather a maximum of 500 best parses, but the limit is not usually reached using feasible beam widths. The mean number of parses per sentence is 176.

4 Results and Discussion

In this section we present results on several reranking methods for joint parsing and semantic role la-
belling. Table 3 compares F1 on the development set of our different reranking methods. The first four rows in Table 3 are baseline systems. We present baselines using gold trees (row 1 in Table 3) and predicted trees (row 2). As shown in previous work, gold trees perform much better than predicted trees.

We also report two cheating baselines to explore the maximum possible performance of a reranking system. First, we report SRL performance of ceiling parse trees (row 3), i.e., if the parse tree from the k-best list is chosen to be closest to the gold tree. This is the best expected performance of a parse reranking approach that maximizes parse F1. Second, we report SRL performance where the parse tree is selected to maximize SRL F1, computing using the gold frame (row 4). There is a significant gap both between parse-F1-reranked trees and SRL-F1-reranked trees, which shows promise for joint reranking. However, the gap between SRL-F1-reranked trees and gold parse trees indicates that reranking of parse lists cannot by itself completely close the gap in SRL performance between gold and predicted parse trees.

4.1 Reranking based on score combination

Equation 1 suggests a straightforward method for reranking: simply pick the parse tree from the k-best list that maximizes \( p(F, t | x) \), in other words, add the log probabilities from the parser and the base SRL system. More generally, we consider weighting the individual probabilities as

\[
s(F, t) = p(F | t, x)^{1 - \alpha} p(t | x)^{\alpha}.
\]

Such a weighted combination is often used in the speech community to combine acoustic and language models.

This reranking method performs poorly, however. No choice of \( \alpha \) performs better than \( \alpha = 1 \), i.e., choosing the 1-best predicted parse tree. Indeed, the more weight given to the SRL score, the worse the combined system performs. The problem is that often a bad parse tree has many nodes which are obviously not constituents: thus \( p(F | t, x) \) for such a bad tree is very high, and therefore not reliable. As more weight is given to the SRL score, the unlabeled recall drops, from 55% when \( \alpha = 0 \) to 71% when \( \alpha = 1 \). Most of the decrease in F1 is due to the drop in unlabeled recall.

4.2 Training a reranker using global features

One potential solution to this problem is to add features of the entire frame, for example, to vote against predicted frames that are missing key arguments. But such features depend globally on the entire frame, and cannot be represented by local classifiers. One way to train these global features is to learn a linear classifier that selects a parse / frame pair from the ranked list, in the manner of Collins (2000). Reranking has previously been applied to semantic role labeling by Toutanova et al. (2005), from which we use several features. The difference between this paper and Toutanova et al. is that instead of reranking k-best SRL frames of a single parse tree, we are reranking 1-best SRL frames from the k-best parse trees.

Because of the the computational expense of training on k-best parse tree lists for each of 30,000 sentences, we train the reranker only on sections 15–18 of the Treebank (the same subset used in previous CoNLL competitions). We train the reranker using LogLoss, rather than the boosting loss used by Collins. We also restrict the reranker to consider only the top 25 parse trees.

This globally-trained reranker uses all of the features from the local model, and the following global features: (a) sequence features, i.e., the linear sequence of argument labels in the sentence (e.g. A0_V_A1), (b) the log probability of the parse tree, (c) has-arg features, that is, for each argument type a binary feature indicating whether it appears in the frame, (d) the conjunction of the predicate and has-arg feature, and (e) the number of nodes in the tree classified as each argument type.

The results of this system on the development set are given in Table 3 (row 6). Although this performs better than the score combination method, it is still no better than simply taking the 1-best parse tree. This may be due to the limited training set we used in the reranking model. A base SRL model trained only on sections 15–18 has 61.26 F1, so in comparison, reranking provides a modest improvement. This system is the one that we submitted as our official submission. The results on the test sets are given in Table 4.

5 Summing over parse trees

In this section, we sketch a different approach to joint SRL and parsing that does not use reranking at all. Maximizing over parse trees can mean that poor parse trees can be selected if their semantic labeling has an erroneously high score. But we are not actually interested in selecting a good parse tree; all we want is a good semantic frame. This means that we should select the semantic frame
6 Conclusion and Related Work

In this paper, we have considered several methods for reranking parse trees using information from semantic role labeling. So far, we have not been able to show improvement over selecting the 1-best parse tree. Gildea and Jurafsky (Gildea and Jurafsky, 2002) also report results on reranking parses using an SRL system, with negative results. In this paper, we confirm these results with a MaxEnt-trained SRL model, and we extend them to show that weighing the probabilities does not help either.

Our results with Collins-style reranking are too preliminary to draw definite conclusions, but the potential improvement does not appear to be great. In future work, we will explore the max-sum approach, which has promise to avoid the pitfalls of max-max reranking approaches.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by National Science Foundation under NSF grants #IIS-0326249 and #IIS-0427594, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References


