Training a Log-Linear Parser with Loss Functions via Softmax-Margin

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:

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.
Training a Log-Linear Parser with Loss Functions via Softmax-Margin

Michael Auli
School of Informatics
University of Edinburgh
m.auli@sms.ed.ac.uk

Adam Lopez
HLTCOE
Johns Hopkins University
alopez@cs.jhu.edu

Abstract

Log-linear parsing models are often trained by optimizing likelihood, but we would prefer to optimise for a task-specific metric like F-measure. Softmax-margin is a convex objective for such models that minimises a bound on expected risk for a given loss function, but its naïve application requires the loss to decompose over the predicted structure, which is not true of F-measure. We use softmax-margin to optimise a log-linear CCG parser for a variety of loss functions, and demonstrate a novel dynamic programming algorithm that enables us to use it with F-measure, leading to substantial gains in accuracy on CCG-Bank. When we embed our loss-trained parser into a larger model that includes supertagging features incorporated via belief propagation, we obtain further improvements and achieve a labelled/unlabelled dependency F-measure of 89.3%/94.0% on gold part-of-speech tags, and 87.2%/92.8% on automatic part-of-speech tags, the best reported results for this task.

1 Introduction

Parsing models based on Conditional Random Fields (CRFs; Lafferty et al., 2001) have been very successful (Clark and Curran, 2007; Finkel et al., 2008). In practice, they are usually trained by maximising the conditional log-likelihood (CLL) of the training data. However, it is widely appreciated that optimizing for task-specific metrics often leads to better performance on those tasks (Goodman, 1996; Och, 2003).

2 Softmax-Margin Training

The softmax-margin objective modifies the standard likelihood objective for CRF training by reweighting
each possible outcome of a training input according to its risk, which is simply the loss incurred on a particular example. This is done by incorporating the loss function directly into the linear scoring function of an individual example.

Formally, we are given \( m \) training pairs \((x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\), where each \( x^{(i)} \in X \) is drawn from the set of possible inputs, and each \( y^{(i)} \in Y(x^{(i)}) \) is drawn from a set of possible instance-specific outputs. We want to learn the \( K \) parameters \( \theta \) of a log-linear model, where each \( \lambda_k \in \theta \) is the weight of an associated feature \( h_k(x, y) \). Function \( f(x, y) \) maps input/output pairs to the vector \( h_1(x, y) \ldots h_K(x, y) \), and our log-linear model assigns probabilities in the usual way.

\[
p(y|x) = \frac{\exp\{\theta^T f(x, y)\}}{\sum_{y' \in Y(x)} \exp\{\theta^T f(x, y')\}} \tag{1}
\]

The conditional log-likelihood objective function is given by Eq. 2 (Figure 1). Now consider a function \( \ell(y, y') \) that returns the loss incurred by choosing to output \( y' \) when the correct output is \( y \). The softmax-margin objective simply modifies the unnormalised, unexponentiated score \( \theta^T f(x, y') \) by adding \( \ell(y, y') \) to it. This yields the objective function (Eq. 3) and gradient computation (Eq. 4) shown in Figure 1.

This straightforward extension has several desirable properties. In addition to having a probabilistic interpretation, it is related to maximum margin and minimum-risk frameworks, it can be shown to minimise a bound on expected risk, and it is convex (Gimpel and Smith, 2010b).

We can also see from Eq. 4 that the only difference from standard CLL training is that we must compute feature expectations with respect to the cost-augmented scoring function. As Gimpel and Smith (2010a) discuss, if the loss function decomposes over the predicted structure, we can treat its decomposed elements as unweighted features that fire on the corresponding structures, and compute expectations in the normal way. In the case of our parser, where we compute expectations using the inside-outside algorithm, a loss function decomposes if it decomposes over spans or productions of a CKY chart.

\section{Loss Functions for Parsing}

Ideally, we would like to optimise our parser towards a task-based evaluation. Our CCG parser is evaluated on labeled, directed dependency recovery using F-measure (Clark and Hockenmaier, 2002). Under this evaluation we will represent output \( y' \) and ground truth \( y \) as variable-sized sets of dependencies. We can then compute precision \( P(y, y') \), recall \( R(y, y') \), and F-measure \( F_1(y, y') \).

\[
P(y, y') = \frac{|y \cap y'|}{|y'|} \tag{5}
\]

\[
R(y, y') = \frac{|y \cap y'|}{|y|} \tag{6}
\]

\[
F_1(y, y') = \frac{2PR}{P + R} = \frac{2|y \cap y'|}{|y| + |y'|} \tag{7}
\]

These metrics are positively correlated with performance – they are gain functions. To incorporate them in the softmax-margin framework we reformulate them as loss functions by subtracting from one.

\subsection{Computing F-Measure-Augmented Expectations at the Sentence Level}

Unfortunately, none of these metrics decompose over parses. However, the individual statistics that are used to compute them do decompose, a fact we will exploit to devise an algorithm that computes the necessary expectations. Note that since \( y \) is fixed, \( F_1 \) is a function of two integers: \( |y \cap y'| \), representing the number of correct dependencies in \( y' \); and \( |y'| \), representing the total number of dependencies in \( y' \), which we will denote as \( n \) and \( d \), respectively.\(^1\) Each pair \( \langle n, d \rangle \) leads to a different value of \( F_1 \). Importantly, both \( n \) and \( d \) decompose over parses.

The key idea will be to treat \( F_1 \) as a non-local feature of the parse, dependent on values \( n \) and \( d \).\(^2\) To compute expectations we split each span in an otherwise usual inside-outside computation by all pairs \( \langle n, d \rangle \) incident at that span.

Formally, our goal will be to compute expectations over the sentence \( a_1 \ldots a_L \). In order to abstract away from the particulars of CCG we present the algorithm in relatively familiar terms as a variant of

\(^1\)For numerator and denominator.

\(^2\)This is essentially the same trick used in the oracle F-measure algorithm of Huang (2008), and indeed our algorithm is a sum-product variant of that max-product algorithm.
the classic inside-outside algorithm (Baker, 1979). We use the notation \( a : A \) for lexical entries and \( BC \Rightarrow A \) to indicate that categories \( B \) and \( C \) combine to form category \( A \) via forward or backward composition or application.\(^3\) The weight of a rule is denoted with \( w \). The classic algorithm associates inside score \( I(A_{i,j}) \) and outside score \( O(A_{i,j}) \) with category \( A \) spanning sentence positions \( i \) through \( j \), computed via the following recursions.

\[
I(A_{i,j+1}) = w(a_{i+1} : A) \\
I(A_{i,j}) = \sum_{k,B,C} I(B_{i,k})I(C_{k,j})w(BC \Rightarrow A) \\
I(GOAL) = I(S_0,L) \\
O(GOAL) = 1 \\
O(A_{i,j}) = \sum_{k,B,C} O(C_{i,k})I(B_{j,k})w(AB \Rightarrow C) + \\
\sum_{k,B,C} O(C_{k,j})I(B_{k,i})w(BA \Rightarrow C)
\]

The expectation of \( A \) spanning positions \( i \) through \( j \) is then \( I(A_{i,j})O(A_{i,j})/I(GOAL) \).

Our algorithm extends these computations to state-split items \( A_{i,j,n,d} \).\(^4\) Using functions \( n(\cdot) \) and \( d(\cdot) \) to respectively represent the number of correct and total dependencies introduced by a parsing action, we present our algorithm in Fig. 3. The final inside equation and initial outside equation incorporate the loss function for all derivations having a particular F-score, enabling us to obtain the desired expectations. A simple modification of the goal equations enables us to optimise precision, recall or a weighted F-measure.

To analyze the complexity of this algorithm, we must ask: how many pairs \( \langle n, d \rangle \) can be incident at each span? A CCG parser does not necessarily return one dependency per span (see Figure 2 for an example), so \( d \) is not necessarily equal to the sentence length \( L \) as it might be in many dependency parsers, though it is still bounded by \( O(L) \). However, this behavior is sufficiently uncommon that we expect all parses of a sentence, good or bad, to have close to \( L \) dependencies, and hence we expect the range of \( d \) to be constant on average. Furthermore, \( n \) will be bounded from below by zero and from above by \( \min(|y|,|y'|) \). Hence the set of all possible F-measures for all possible parses is bounded by \( O(L^2) \), but on average it should be closer to \( O(L) \). Following McAllester (1999), we can see from inspection of the free variables in Fig. 3 that the algorithm requires worst-case \( O(L^7) \) and average-case \( O(L^5) \) time complexity, and worse-case \( O(L^4) \) and average-case \( O(L^3) \) space complexity.

Note finally that while this algorithm computes exact sentence-level expectations, it is approximate at the corpus level, since F-measure does not decompose over sentences. We give the extension to exact corpus-level expectations in Appendix A.

### 3.2 Approximate Loss Functions

We will also consider approximate but more efficient alternatives to our exact algorithms. The idea is to use cost functions which only utilise statistics

\[
\min_\theta \sum_{i=1}^m \left[ -\theta^T f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^T f(x^{(i)}, y)\} \right]
\]

\[
\min_\theta \sum_{i=1}^m \left[ -\theta^T f(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\theta^T f(x^{(i)}, y) + \ell(y^{(i)}, y)\} \right]
\]

\[
\frac{\partial}{\partial \lambda_k} = \sum_{i=1}^m \left[ -h_k(x^{(i)}, y^{(i)}) + \sum_{y \in \mathcal{Y}(x^{(i)})} \frac{\exp\{\theta^T f(x^{(i)}, y) + \ell(y^{(i)}, y)\}}{\sum_{y' \in \mathcal{Y}(x^{(i)})} \exp\{\theta^T f(x^{(i)}, y') + \ell(y^{(i)}, y')\}} h_k(x^{(i)}, y) \right]
\]
\[ I(A_{i,i+1,n,d}) = w(a_{i+1} : A) \text{ iff } n = n_+(a_{i+1} : A), d = d_+(a_{i+1} : A) \]
\[ I(A_{i,j,n,d}) = \sum_{k,B,C} \sum_{\{n',n' : n' + n_+(BC\Rightarrow A) = n\}} \sum_{\{d',d'' : d' + d'' + d_+(BC\Rightarrow A) = d\}} I(B_{i,k,n',d'}) I(C_{k,j,n'',d''}) w(BC \Rightarrow A) \]
\[ I(GOAL) = \sum_{n,d} I(S_0,L,n,d) \left( 1 - \frac{2n}{d + |y|} \right) \]
\[ O(S_0,n,d) = \left( 1 - \frac{2n}{d + |y|} \right) \]
\[ O(A_{i,j,n,d}) = \sum_{k,B,C} \sum_{\{n',n' : n' - n_+(AB\Rightarrow C) = n\}} \sum_{\{d',d'' : d' - d'' - d_+(AB\Rightarrow C) = d\}} O(C_{i,k,n',d'}) I(B_{j,k,n'',d''}) w(AB \Rightarrow C) + \]
\[ \sum_{k,B,C} \sum_{\{n',n' : n' - n_+(BA\Rightarrow C) = n\}} \sum_{\{d',d'' : d' - d'' - d_+(BA\Rightarrow C) = d\}} O(C_{k,j,n',d'}) I(B_{k,i,n'',d''}) w(BA \Rightarrow C) \]

Figure 3: State-split inside and outside recursions for computing softmax-margin with F-measure.

Figure 2: Example of flexible dependency realisation in CCG: Our parser (Clark and Curran, 2007) creates dependencies arising from coordination once all conjuncts are found and treats “and” as the syntactic head of coordinations. The coordination rule (Φ) does not yet establish the dependency “and - pears” (dotted line); it is the backward application (<) in the larger span, “apples and pears”, that establishes this, together with “and - pears”. CCG also deals with unbounded dependencies which potentially lead to more dependencies than words (Steedman, 2000); in this example a unification mechanism creates the dependencies “likes - apples” and “likes - pears” in the forward application (>). For further examples and a more detailed explanation of the mechanism as used in the C&C parser refer to Clark et al. (2002).

available within the current local structure, similar to those used by Taskar et al. (2004) for tracking constituent errors in a context-free parser. We design three simple losses to approximate precision, recall and F-measure on CCG dependency structures.

Let \( T(y) \) be the set of parsing actions required to build parse \( y \). Our decomposable approximation to precision simply counts the number of incorrect dependencies using the local dependency counts, \( n_+() \) and \( d_+() \).

\[ DecP(y) = \sum_{t \in T(y)} d_+(t) - n_+(t) \] (8)

To compute our approximation to recall we require the number of gold dependencies, \( c_+() \), which should have been introduced by a particular parsing action. A gold dependency is due to be recovered by a parsing action if its head lies within one child span and its dependent within the other. This yields a decomposed approximation to recall that counts the number of missed dependencies.

\[ DecR(y) = \sum_{t \in T(y)} c_+(t) - n_+(t) \] (9)
Unfortunately, the flexible handling of dependencies in CCG complicates our formulation of $c_+$, rendering it slightly more approximate. The unification mechanism of CCG sometimes causes dependencies to be realised later in the derivation, at a point when both the head and the dependent are in the same span, violating the assumption used to compute $c_+$ (see again Figure 2). Exceptions like this can cause mismatches between $n_+$ and $c_+$. We set $c_+ = n_+$ whenever $c_+ < n_+$ to account for these occasional discrepancies.

Finally, we obtain a decomposable approximation to F-measure.

$$ DecF_1(y) = DecP(y) + DecR(y) \quad (10) $$

4 Experiments

Parsing Strategy. CCG parsers use a pipeline strategy: we first multitag each word of the sentence with a small subset of its possible lexical categories using a supertagger, a sequence model over these categories (Bangalore and Joshi, 1999; Clark, 2002). Then we parse the sentence under the requirement that the lexical categories are fixed to those preferred by the supertagger. In our experiments we used two variants on this strategy.

First is the adaptive supertagging (AST) approach of Clark and Curran (2004). It is based on a step function over supertagger beam widths, relaxing the pruning threshold for lexical categories only if the parser fails to find an analysis. The process either succeeds and returns a parse after some iteration or gives up after a predefined number of iterations. As Clark and Curran (2004) show, most sentences can be parsed with very tight beams.

Reverse adaptive supertagging is a much less aggressive method that seeks only to make sentences parsable when they otherwise would not be due to an impractically large search space. Reverse AST starts with a wide beam, narrowing it at each iteration only if a maximum chart size is exceeded. Table 1 shows beam settings for both strategies.

Adaptive supertagging aims for speed via pruning while the reverse strategy aims for accuracy by exposing the parser to a larger search space. Although Clark and Curran (2007) found no actual improvements from the latter strategy, we will show that with our softmax-margin-trained models it can have a substantial effect.

Parser. We use the C&C parser (Clark and Curran, 2007) and its supertagger (Clark, 2002). Our baseline is the hybrid model of Clark and Curran (2007), which contains features over both normal-form derivations and CCG dependencies. The parser relies solely on the supertagger for pruning, using exact CKY for search over the pruned space. Training requires calculation of feature expectations over packed charts of derivations. For training, we limited the number of items in this chart to 0.3 million, and for testing, 1 million. We also used a more permissive training supertagger beam (Table 2) than in previous work (Clark and Curran, 2007). Models were trained with the parser’s L-BFGS trainer.

Evaluation. We evaluated on CCGbank (Hockenmaier and Steedman, 2007), a right-most normal-form CCG version of the Penn Treebank. We use sections 02-21 (39603 sentences) for training, section 00 (1913 sentences) for development and section 23 (2407 sentences) for testing. We supply gold-standard part-of-speech tags to the parsers. We evaluate on labelled and unlabelled predicate argument structure recovery and supertag accuracy.

4.1 Training with Maximum F-measure Parses

So far we discussed how to optimise towards task-specific metrics via changing the training objective. In our first experiment we change the data on which we optimise CLL. This is a kind of simple baseline to our later experiments, attempting to achieve the same effect by simpler means. Specifically, we use the algorithm of Huang (2008) to generate oracle F-measure parses for each sentence. Updating towards these oracle parses corrects the reachability problem in standard CLL training. Since the supertagger is used to prune the training forests, the correct parse is sometimes pruned away – reducing data utilisation to 91%. Clark and Curran (2007) correct for this by adding the gold tags to the parser input. While this increases data utilisation, it biases the model by training in an idealised setting not available at test time. Using oracle parses corrects this bias while permitting 99% data utilisation. The labelled F-score of the oracle parses lies at 98.1%. Though we expected that this might result in some improvement, results (Table 3) show that this has no
Table 1: Beam step function used for standard (AST) and less aggressive (Reverse) AST throughout our experiments. Parameter $\beta$ is a beam threshold while $k$ bounds the number of lexical categories considered for each word.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Parameter</th>
<th>Iteration 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AST</td>
<td>$\beta$ (beam width)</td>
<td>0.075</td>
<td>0.03</td>
<td>0.01</td>
<td>0.005</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>$k$ (dictionary cutoff)</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>150</td>
</tr>
<tr>
<td>Reverse</td>
<td>$\beta$</td>
<td>0.001</td>
<td>0.005</td>
<td>0.01</td>
<td>0.03</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>$k$</td>
<td>150</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 2: Beam step functions used for training: The first row shows the large scale settings used for most experiments and the standard C&C settings. (cf. Table 1)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Parameter</th>
<th>Iteration 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>$\beta$</td>
<td>0.001</td>
<td>0.001</td>
<td>0.0045</td>
<td>0.0055</td>
<td>0.01</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>$k$</td>
<td>150</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>C&amp;C '07</td>
<td>$\beta$</td>
<td>0.0045</td>
<td>0.0055</td>
<td>0.01</td>
<td>0.05</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$k$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Performance on section 00 of CCGbank when comparing models trained with treebank-parses (Baseline) and maximum F-score parses (Max-F) using adaptive supertagging as well as a combination of CCGbank and Max-F parses. Evaluation is based on labelled and unlabelled F-measure (LF/UF), precision (LP/UP) and recall (LR/UR).

<table>
<thead>
<tr>
<th></th>
<th>LF</th>
<th>LP</th>
<th>LR</th>
<th>UF</th>
<th>UP</th>
<th>UR</th>
<th>Data Util (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>87.40</td>
<td>87.85</td>
<td>86.95</td>
<td>93.11</td>
<td>93.59</td>
<td>92.63</td>
<td>91%</td>
</tr>
<tr>
<td>Max-F Parsed</td>
<td>87.46</td>
<td>87.95</td>
<td>86.98</td>
<td>93.09</td>
<td>93.61</td>
<td>92.57</td>
<td>99%</td>
</tr>
<tr>
<td>CCGbank+Max-F</td>
<td>87.45</td>
<td>87.96</td>
<td>86.94</td>
<td>93.09</td>
<td>93.63</td>
<td>92.55</td>
<td>99%</td>
</tr>
</tbody>
</table>

4.2 Training with the Exact Algorithm

We first tested our assumptions about the feasibility of training with our exact algorithm by measuring the amount of state-splitting. Figure 4 plots the average number of splits per span against the relative span-frequency; this is based on a typical set of training forests containing over 600 million states. The number of splits increases exponentially with span size but equally so decreases the number of spans with many splits. Hence the small number of states with a high number of splits is balanced by a large number of spans with only a few splits: The highest number of splits per span observed with our settings was 4888 but we find that the average number of splits lies at 44. Encouragingly, this enables experimentation in all but very large scale settings.

Figure 5 shows the distribution of $n$ and $d$ pairs across all split-states in the training corpus; since $n$, the number of correct dependencies, over $d$, the number of all recovered dependencies, is precision, the graph shows that only a minority of states have either very high or very low precision. The range of values suggests that the softmax-margin criterion...
will have an opportunity to substantially modify the expectations, hopefully to good effect.

We next turn to the question of optimization with these algorithms. Due to the significant computational requirements, we used the computationally less intensive normal-form model of Clark and Curran (2007) as well as their more restrictive training beam settings (Table 2). We train on all sentences of the training set as above and test with AST. In order to provide greater control over the influence of the loss function, we introduce a multiplier $\tau$, which simply amends the second term of the objective function (3) to:

$$\log \sum_{y \in Y(x^i)} \exp\{g^T f(x^i, y) + \tau \times \ell(y^i, y)\}$$

Figure 6 plots performance of the exact loss functions across different settings of $\tau$ on various evaluation criteria, for models restricted to at most 3000 items per chart at training time to allow rapid experimentation with a wide parameter set. Even in this constrained setting, it is encouraging to see that each loss function performs best on the criteria it optimises. The precision-trained parser also does very well on F-measure; this is because the parser has a tendency to perform better in terms of precision than recall.

### 4.3 Exact vs. Approximate Loss Functions

With these results in mind, we conducted a comparison of parsers trained using our exact and approximate loss functions. Table 4 compares their performance head to head when restricting training chart sizes to 100,000 items per sentence, the largest setting our computing resources allowed us to experiment with. The results confirm that the loss-trained models improve over a likelihood-trained baseline, and furthermore that the exact loss functions seem to have the best performance. However, the approximations are extremely competitive with their exact counterparts. Because they are also efficient, this makes them attractive for larger-scale experiments. Training time increases by an order of magnitude with exact loss functions despite increased theoretical complexity (§3.1); there is no significant change with approximate loss functions.

Table 5 shows performance of the approximate losses with the large scale settings initially outlined (§4). One striking result is that the softmax-margin trained models coax more accurate parses from the larger search space, in contrast to the likelihood-trained models. Our best loss model improves the labelled F-measure by over 0.8%.

### 4.4 Combination with Integrated Parsing and Supertagging

As a final experiment, we embed our loss-trained model into an integrated model that incorporates Markov features over supertags into the parsing model (Auli and Lopez, 2011). These features have serious implications on search: even allowing for the observation of Fowler and Penn (2010) that our CCG is weakly context-free, the search problem is equivalent to finding the optimal derivation in the weighted intersection of a regular and context-free language (Bar-Hillel et al., 1964), making search very expensive. Therefore parsing with this model requires approximations.

To experiment with this combined model we use loopy belief propagation (LBP; Pearl et al., 1985), previously applied to dependency parsing by Smith and Eisner (2008). A more detailed account of its application to our combined model can be found in (2011), but we sketch the idea here. We construct a graphical model with two factors: one is a distribu-
Figure 6: Performance of exact cost functions optimizing F-measure, precision and recall in terms of (a) labelled F-measure, (b) precision, (c) recall and (d) supertag accuracy across various settings of $\tau$ on the development set.

Table 4: Performance of exact and approximate loss functions against conditional log-likelihood (CLL): decomposable precision (DecP), recall (DecR) and F-measure (DecF1) versus exact precision (P), recall (R) and F-measure (F1). Evaluation is based on labelled and unlabelled F-measure (LF/UF), precision (LP/UP) and recall (LR/UR).
Table 5: Performance of decomposed loss functions in large-scale training setting. Evaluation is based on labelled and unlabelled F-measure (LF/UF) and supertag accuracy (ST).

<table>
<thead>
<tr>
<th></th>
<th>AST</th>
<th></th>
<th>Reverse</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LF</td>
<td>UF</td>
<td>ST</td>
<td>LF</td>
<td>UF</td>
<td>ST</td>
<td>LF</td>
<td>UF</td>
<td>ST</td>
<td>LF</td>
<td>UF</td>
<td>ST</td>
<td></td>
</tr>
<tr>
<td>CLL</td>
<td>87.38</td>
<td>93.08</td>
<td>94.21</td>
<td>87.36</td>
<td>93.13</td>
<td>93.99</td>
<td>87.73</td>
<td>93.09</td>
<td>94.33</td>
<td>87.65</td>
<td>93.06</td>
<td>94.01</td>
<td></td>
</tr>
<tr>
<td>DecP</td>
<td>87.35</td>
<td>92.99</td>
<td>94.25</td>
<td>87.75</td>
<td>93.25</td>
<td>94.22</td>
<td>88.10</td>
<td>93.26</td>
<td>94.51</td>
<td>88.51</td>
<td>93.50</td>
<td>94.39</td>
<td></td>
</tr>
<tr>
<td>DecR</td>
<td>87.48</td>
<td>93.00</td>
<td>94.34</td>
<td>87.70</td>
<td>93.16</td>
<td>94.30</td>
<td>87.66</td>
<td>92.83</td>
<td>94.38</td>
<td>87.77</td>
<td>92.91</td>
<td>94.22</td>
<td></td>
</tr>
<tr>
<td>DecF1</td>
<td><strong>87.67</strong></td>
<td><strong>93.23</strong></td>
<td><strong>94.39</strong></td>
<td><strong>88.12</strong></td>
<td><strong>93.52</strong></td>
<td><strong>94.46</strong></td>
<td>88.09</td>
<td><strong>93.28</strong></td>
<td><strong>94.50</strong></td>
<td><strong>88.58</strong></td>
<td><strong>93.57</strong></td>
<td><strong>94.53</strong></td>
<td></td>
</tr>
</tbody>
</table>

5 Conclusion and Future Work

The softmax-margin criterion is a simple and effective approach to training log-linear parsers. We have shown that it is possible to compute exact sentence-level losses under standard parsing metrics, not only approximations (Taskar et al., 2004). This enables us to show the effectiveness of these approximations, and it turns out that they are excellent substitutes for exact loss functions. Indeed, the approximate losses are as easy to use as standard conditional log-likelihood.

Empirically, softmax-margin training improves parsing performance across the board, beating the state-of-the-art CCG parsing model of Clark and Curran (2007) by up to 0.8% labelled F-measure. It also proves robust, improving a stronger baseline based on a combined parsing and supertagging model. Our final result of 89.3%/94.0% labelled and unlabelled F-measure is the best result reported for CCG parsing accuracy, beating the original C&C baseline by up to 1.5%.

In future work we plan to scale our exact loss functions to larger settings and to explore training with loss functions within loopy belief propagation. Although we have focused on CCG parsing in this work, we expect our methods to be equally applicable to parsing with other grammar formalisms including context-free grammar or LTAG.

Acknowledgements

We would like to thank Stephen Clark, Christos Christodoulopoulos, Mark Granroth-Wilding, Gholamreza Haffari, Alexandre Klementiev, Tom Kwiatkowski, Kira Mourao, Matt Post, and Mark Steedman for helpful discussion related to this work and comments on previous drafts, and the
Table 6: Performance of combined parsing and supertagging with belief propagation (BP); using decomposed-F1 as parser-loss function and supertag-accuracy (SA) as loss in the supertagger.

<table>
<thead>
<tr>
<th></th>
<th>section 00 (dev)</th>
<th>section 23 (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AST Reverse</td>
<td>AST Reverse</td>
</tr>
<tr>
<td></td>
<td>LF   UF  ST</td>
<td>LF   UF  ST</td>
</tr>
<tr>
<td>CLL</td>
<td>87.38 93.08 94.21</td>
<td>87.36 93.13 93.99</td>
</tr>
<tr>
<td>BP</td>
<td>87.67 93.26 94.43</td>
<td>88.25 93.33 94.60</td>
</tr>
<tr>
<td>+DecF1</td>
<td>87.90 93.40 94.52</td>
<td>88.32 93.32 94.66</td>
</tr>
<tr>
<td>+SA</td>
<td>87.73 93.28 94.49</td>
<td>87.65 93.06 94.01</td>
</tr>
</tbody>
</table>

Table 7: Results on automatically assigned POS tags. Petrov I-5 is based on the parser output of Fowler and Penn (2010); evaluation is based on sentences for which all parsers returned an analysis.

<table>
<thead>
<tr>
<th></th>
<th>section 00 (dev)</th>
<th>section 23 (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LF   LP  LR  UF  UP  UR</td>
<td>LF   LP  LR  UF  UP  UR</td>
</tr>
<tr>
<td>CLL</td>
<td>85.53 85.73 85.33 91.99 92.20 91.77</td>
<td>85.74 85.90 85.58 91.92 92.09 91.75</td>
</tr>
<tr>
<td>Petrov</td>
<td>85.79 86.09 85.50 92.44 92.76 92.13</td>
<td>86.01 86.29 85.73 92.34 92.64 92.04</td>
</tr>
<tr>
<td>BP</td>
<td>86.45 86.75 86.17 92.60 92.92 92.29</td>
<td>86.84 87.08 86.61 92.57 92.82 92.32</td>
</tr>
<tr>
<td>+DecF1</td>
<td>86.73 87.07 86.39 92.79 93.16 92.43</td>
<td>87.08 87.37 86.78 92.68 93.00 92.37</td>
</tr>
<tr>
<td>+SA</td>
<td>86.51 86.86 86.16 92.60 92.98 92.23</td>
<td>87.20 87.50 86.90 92.76 93.08 92.44</td>
</tr>
</tbody>
</table>

Anonymous reviewers for helpful comments. We also acknowledge funding from EPSRC grant EP/P504171/1 (Auli); and the resources provided by the Edinburgh Compute and Data Facility.

A Computing F-Measure-Augmented Expectations at the Corpus Level

To compute exact corpus-level expectations for softmax-margin using F-measure, we add an additional transition before reaching the GOAL item in our original program. To reach it, we must parse every sentence in the corpus, associating statistics of aggregate \( \langle n, d \rangle \) pairs for the entire training set in intermediate symbols \( \Gamma^{(1)} \ldots \Gamma^{(m)} \) with the following inside recursions.

\[
I(\Gamma^{(1)}_{n,d}) = I(S^{(1)}_{0,|x^{(1)}|,n,d})
\]
\[
I(\Gamma^{(f)}_{n,d}) = \sum_{n',n'';n''+n'=n} I(\Gamma^{(f-1)}_{n',d'}) I(S^{(f)}_{0,N,N''},d')
\]
\[
I(GOAL) = \sum_{n,d} I(\Gamma^{(m)}_{n,d}) \left( 1 - \frac{2n}{d + |y|} \right)
\]

Outside recursions follow straightforwardly. Implementation of this algorithm would require substantial distributed computation or external data structures, so we did not attempt it.

References


