Re-estimation of Lexical Parameters for Treebank PCFGs

Tejaswini Deoskar
Department of Linguistics
Cornell University
td72@cornell.edu

Abstract

We present procedures which pool lexical information estimated from unlabeled data via the Inside-Outside algorithm, with lexical information from a treebank PCFG. The procedures produce substantial improvements (up to 31.6% error reduction) on the task of determining subcategorization frames of novel verbs, relative to a smoothed Penn Treebank-trained PCFG. Even with relatively small quantities of unlabeled training data, the re-estimated models show promising improvements in labeled bracketing *f*-scores on Wall Street Journal parsing, and substantial benefit in acquiring the subcategorization preferences of low-frequency verbs.

1 Introduction

In order to obtain the meaning of a sentence automatically, it is necessary to have access to its syntactic analysis at some level of complexity. Many NLP applications like translation, question-answering, etc. might benefit from the availability of syntactic parses. Probabilistic parsers trained over labeled data have high accuracy on in-domain data: lexicalized parsers get an *f*-score of up to 90.0% on Wall Street Journal data (Charniak and Johnson (2005)'s re-ranking parser), while recently, unlexicalized PCFGs have also been shown to perform much better than previously believed (Klein and Manning, 2003). However, the limited size of annotated training data results in many parameters of a PCFG being badly estimated when trained on annotated data. The Zipfian nature of a text corpus results in PCFG parameters related to the properties of specific words being especially badly estimated. For instance, about 38% of verbs in the training sections of the Penn Treebank (PTB) (Marcus et al., 1993) occur only once – the lexical properties of these verbs (such as their most common subcategorization frames) cannot be represented accurately in a model trained exclusively on the Penn Treebank.

The research reported here addresses this issue. We start with an unlexicalized PCFG trained on the PTB. We then re-estimate the parameters of this PCFG from raw text using an unsupervised estimation method based on the Inside-Outside algorithm (Lari and Young, 1990), an instance of the Expectation Maximization algorithm (Dempster et al., 1977) for PCFG induction. The re-estimation improves *f*-score on the standard test section of the PTB significantly. Our focus is on learning lexical parameters i.e. those parameters related to the lexico-syntactic properties of open-class words. Examples of such properties are: subcategorization frames of verbs and nouns, attachment preference of adverbs to sentential, verbal or nominal nodes, attachment preference of PPs to a verbal or nominal node, etc.

The current research is related to semi-supervised training paradigms like self-training – these methods are currently being explored to improve the performance of existing PCFG models by utilizing unlabeled data. For example, McCloskey et al. (2006) achieve a 1.1% improvement in labeled bracketing *f*-score by the use of unlabeled data to self-train the parser-reranker system from Charniak and Johnson (2005). Earlier research on inside-outside estimation of PCFG models has reported some positive results as well

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 193–200
Manchester, August 2008

(Pereira and Schabes, 1992; Carroll and Rooth, 1998; Beil et al., 1999; imWalde, 2002). In some of these cases, an initial model is derived by other means – inside-outside is used to *re-estimate* the initial model. However, many questions still remain open about its efficacy for PCFG re-estimation. Grammars used previously have not been treebank grammars (for e.g., Carroll and Rooth (1998) and Beil et al. (1999) used hand-crafted grammars), hence these models could not be evaluated according to standardized evaluations in the parsing literature. In the current work, we use a Penn Treebank based grammar; hence all re-estimated grammars can be evaluated using standardized criteria.

The rest of the paper is organized as follows: First, we describe in brief the construction of an unlexicalized PCFG from the PTB. We then describe a procedure based on the inside-outside algorithm to re-estimate the lexical parameters of this PCFG from unlabeled Wall Street Journal data. Finally, we present evaluations of the re-estimated models, based on labeled bracketing measures and on the detection of subcategorization frames of verbs: there is a 31.6% reduction in error for novel verbs and up to 8.97% reduction in overall subcategorization error.

2 Unlexicalized treebank PCFG

We build an unlexicalized PCFG from the standard training sections of the PTB. As is common (Collins, 1997; Johnson, 1998; Klein and Manning, 2003; Schmid, 2006), the treebank is first transformed in various ways, in order to give an accurate PCFG. In our framework, treebank trees are augmented with extra features; the methodology involves constructing a feature-constraint grammar from a context-free treebank backbone grammar. The detailed methodology is described in Deoskar and Rooth (2008)[1]. A PCFG is trained on the transformed treebank, with these added features incorporated into the PCFG's non-terminal categories. The framework affords us the flexibility to stipulate the features to be incorporated in the PCFG categories, as parameters of the PCFG.

Our features are largely designed to have a linguistically relevant interpretation[2]. For example, there is a feature on verbs which denotes the subcategorization frame of the verb (with values like intransitive, transitive, etc.). Similarly, there are features which denote the type of clause (finite, infinite, small clause, etc.), the subject type of clausal nodes, the attachment of adverbs, valence of nouns, etc. Unlike most existing treebank PCFGs, all PTB function tags are retained, as are all empty categories.

	this paper	Schmid	K&M
Recall	86.5	86.3	85.1
Precision	86.7	86.9	86.3
F-score	86.6	86.6	85.7

Table 1: Labeled bracketing scores, PTB sec. 23.

As a measure of the quality of the transformed-PTB based PCFG, Table 1 gives the labeled bracketing scores on the standard test section 23 of the PTB, comparing them to unlexicalized PCFG scores in (Schmid, 2006) and (Klein and Manning, 2003) (K&M). The current PCFG *f*-score is comparable to the state-of-the-art in unlexicalized PCFGs ((Schmid, 2006), to our knowledge). We stopped grammar development when the *f*-score reached state-of-the-art since our goal was to use this grammar as the initial model and baseline for the unsupervised re-estimation procedure, described in the next section.

3 Inside-Outside Re-estimation

As a basic unsupervised estimation method, we use standard inside-outside estimation of PCFGs, which realizes EM estimation (Lari and Young, 1990; Pereira and Schabes, 1992). We use the notation $I(C, e)$ to designate the new frequency model, computed via inside-outside from the corpus C by using a probability model based on the frequency model e[3]. The iterative inside-outside re-estimation procedure has the following simple form (Eq.1), where each successive frequency model e_{i+1} is estimated from the corpus C using a probability model determined by the previous frequency model e_i. Our notation always refers to fre-

[1]The reason for using this framework (as opposed to using available unlexicalized PCFGs) is that it allows us flexibility in designing features of interest, and can also be used for languages other than English with existing treebanks.

[2]In addition we also have some features that do not have a linguistic interpretation, but result in a good PCFG, such as a *parent* feature on some categories, following Johnson (1998).

[3]The inside-outside algorithm uses an existing grammar model and a raw text corpus (incomplete data) to obtain corresponding complete data (a set of analyses/parses for the corpus sentences). A new grammar model is then estimated from this complete data. See (Prescher, 2003) for an explanation using the standard EM notions of incomplete/complete data.

quency models such as e_i, rather than the relative-frequency probability models they determine[4].

$$\begin{aligned} e_1 &= I(C, e_0) \\ &\ldots \\ e_{i+1} &= I(C, e_i) \end{aligned} \tag{1}$$

3.1 Interleaved Inside-Outside

It is well-known that while lexicalization is useful, lexical parameters determined from the treebank are poorly estimated because of the sparseness of treebank data for particular words (e.g. Hindle and Rooth (1993)). Gildea (2001) and Bikel (2004) show that removing bilexical dependencies hardly hurts the performance of the Collins Model2 parser, although there is the benefit of lexicalization in the form of lexico-syntactic dependencies – structures being conditioned on words. On the other hand, structural parameters are comparatively well-estimated from treebanks since they are not keyed to particular words. Thus, it might be beneficial to use a combination of supervised and unsupervised estimation for lexical parameters, while obtaining syntactic (structural) parameters solely by supervised estimation (i.e. from a treebank). The experiments in this paper are based on this idea. In an unlexicalised PCFG like the one described in §2, it is easy to make the distinction between structural parameters (non-terminal rules) and lexical parameters (preterminal to terminal rules).

To this end, we define a modified inside-outside procedure in which a frequency transformation $T(c, t)$ is interleaved between the iterations of the standard inside-outside procedure. The form of this interleaved procedure is shown in Eq. 2. In Eq. 2, t designates a smoothed treebank model (the smoothing procedure is described later in §3.1.1). This smoothed treebank model is used as the prior model for the inside-outside re-estimation procedure. For each iteration i, c_i represent models obtained by inside-outside estimation. d_i represent *derived* models obtained by performing a transformation T on c_i. The transformation T combines the re-estimated model c_i and the smoothed treebank model t (hence represented as $T(c_i, t)$).

$$\begin{aligned} d_0 &= t && \textit{smoothed treebank model} \\ c_1 &= I(C, d_0) && \textit{estimation step} \\ d_1 &= T(c_1, t) && \textit{transformation step} \\ &\ldots \\ c_{i+1} &= I(C, d_i) && \textit{estimation step} \\ d_{i+1} &= T(c_{i+1}, t) && \textit{transformation step} \end{aligned} \tag{2}$$

The lexical parameters for the treebank model t or the re-estimated models c_i are represented as $t(w, \tau, \iota)$ or $c_i(w, \tau, \iota)$, where w is the terminal word, τ is the PTB-style PoS tag, and ι is the sequence of additional features incorporated into the PoS tag (the entries in our lexicon have the form $w.\tau.\iota$ with an associated frequency). The transformation T preserves the marginal frequencies seen in the treebank model. A marginal tag-incorporation frequency is defined by summation:

$$f(\tau, \iota) = \sum_w f(w, \tau, \iota). \tag{3}$$

The transformation T is used to obtain the derived models d_i and consists of two parts, corresponding to the syntactic and the lexical parameters of d_i:

- The syntactic parameters of d_i are copied from t.
- To obtain the lexical parameters of d_i, lexical parameters from the treebank model t and lexical parameters from the re-estimated model are linearly combined, shown in Eq. 4.

$$d_i(w, \tau, \iota) = (1 - \lambda_{\tau,\iota}) t(w, \tau, \iota) + \lambda_{\tau,\iota} \bar{c_i}(w, \tau, \iota) \tag{4}$$

where $\lambda_{\tau,\iota}$ is a parameter with $0 < \lambda_{\tau,\iota} < 1$ which may depend on the tag and incorporation. The term $\bar{c_i}(w, \tau, \iota)$ in Eq. 4 is obtained by scaling the frequencies in $c_i(w, \tau, \iota)$, as shown in Eq. 5.

$$\bar{c_i}(w, \tau, \iota) = \frac{t(\tau, \iota)}{c_i(\tau, \iota)} c_i(w, \tau, \iota). \tag{5}$$

In terms of probability models determined from the frequency models, the effect of T is to allocate a fixed proportion of the probability mass for each τ, ι to the corpus, and share it out among words w in proportion to relative frequencies $\frac{c_i(w,\tau,\iota)}{c_i(\tau,\iota)}$ in the inside-outside estimate c_i. Eqs. 6 and 7 verify that marginals are preserved in the derived model d.

$$\begin{aligned} \bar{c}(\tau, \iota) &= \textstyle\sum_w \bar{c}(w, \tau, \iota) = \sum_w \frac{t(\tau,\iota)}{c(\tau,\iota)} c(w, \tau, \iota) \\ &= \tfrac{t(\tau,\iota)}{c(\tau,\iota)} \textstyle\sum_w c(w, \tau, \iota) \\ &= \tfrac{t(\tau,\iota)}{c(\tau,i)} c(\tau, \iota) = t(\tau, \iota). \end{aligned} \tag{6}$$

[4] We use a frequency-based notation because we use out-of-the-box software Bitpar (Schmid, 2004) which implements inside-outside estimation – Bitpar reads in frequency models and converts them to relative frequency models. We justify the use of the frequency-based notation by ensuring that all marginal frequencies in the treebank model are always preserved in all other models.

$$
\begin{aligned}
d(\tau,\iota) &= \textstyle\sum_w d(w,\tau,\iota) \\
&= \textstyle\sum_w (1-\lambda_{\tau,\iota}) t(w,\tau,\iota) + \lambda_{\tau,\iota}\bar{c}(w,\tau,\iota) \\
&= (1-\lambda_{\tau,\iota}) \textstyle\sum_w t(w,\tau,\iota) \\
&+ \ \lambda_{\tau,\iota} \textstyle\sum_w \bar{c}(w,\tau,\iota) \\
&= (1-\lambda_{\tau,\iota}) t(\tau,\iota) + \lambda_{\tau,\iota}\bar{c}(\tau,\iota) \\
&= (1-\lambda_{\tau,\iota}) t(\tau,\iota) + \lambda_{\tau,\iota} t(\tau,\iota) \\
&= t(\tau,\iota).
\end{aligned}
\tag{7}
$$

3.1.1 Smoothing the treebank model

To initialize the iterative procedures, a smoothing scheme is required which allocates frequency to combinations of words w and PoS tags τ which are not present in the treebank model but are present in the corpus, and also to all possible incorporations of a PoS tag. Otherwise, if the unsmoothed treebank model (t_0) has zero frequency for some lexical parameter, the inside-outside estimate $I(C, t_0)$ for that parameter would also be zero, and new lexical entries would never be induced.

The smoothed treebank model t is obtained from the unsmoothed model t_0 as follows. First a PoS tagger (Treetagger, (Schmid, 1994)) is run on the unsupervised corpus C, which assigns PTB-style PoS tags to the corpus. Tokens of words and PoS tags are tabulated to obtain a frequency table $g(w,\tau)$. Each frequency $g(w,\tau)$ is split among possible incorporations ι in proportion to a ratio of marginal frequencies in t_0:

$$
g(w,\tau,\iota) = \frac{t_0(\tau,\iota)}{t_0(\tau)} g(w,\tau) \tag{8}
$$

The smoothed model t is defined as an interpolation of g and t_0 for lexical parameters as shown in 9, with syntactic parameters copied from t_0.

$$
t(w,\tau,\iota) = (1-\lambda_{\tau,\iota}) t_0(w,\tau,\iota) + \lambda_{\tau,\iota} g(w,\tau,\iota) \tag{9}
$$

3.2 Experimental setup

The treebank grammar is trained over sections 0-22 of the transformed PTB (minus about 7000 sentences held out for testing). Testset I contains 1331 sentences and is constructed as follows: First, we select 117 verbs whose frequency in PTB sections 0-22 is between 10-20 (mid-frequency verbs). All sentences containing occurrences of these verbs are held out from the training data to form Testset I. The effect of holding out these sentences is to make these 117 verbs *novel* (i.e. unseen in training). This testset is used to evaluate the learning of subcategorization frames of novel verbs. We also construct another testset (Testset II) by holding out every 10^{th} sentence in PTB sections 0-22 (4310 sentences).

The corpus used for re-estimation is about 4 million words of unannotated Wall Street Journal text (year 1997) (sentence length<25 words). The re-estimation was carried out using Bitpar (Schmid, 2004) for inside-outside estimation. The parameter λ in Eq. 4 was set to 0.5 for all τ and ι, giving equal weight to the treebank and the re-estimated lexicons. Starting from a smoothed treebank grammar t, we separately ran 6 iterations of the interleaved estimation procedure defined in Eq. 2, and 4 iterations of standard inside-outside estimation. This gave us two series of models corresponding to the two procedures.

4 Labeled Bracketing Results

As a basic evaluation of the re-estimated grammars, we report the labeled bracketing scores on the standard test section 23 of the PTB (Table 2). Using the re-estimated models, maximum probability (viterbi) parses were obtained for all sentences in sec. 23, after stripping away the treebank annotation, including the pre-terminal tag. The baseline is the treebank model t_{0t}[5]. The scores for re-estimated grammars from successive iterations are under columns It 1, It 2, etc. All models obtained using the interleaved procedure show an improvement over the baseline. The best model is obtained after 2 iterations, after which the score reduces a little. Statistically significant improvements are marked with *, with p<0.005 for recall and p<0.0001 for precision for the best model. Table 2 also shows scores for grammars estimated using the standard inside-outside procedure. The first re-estimated model is better than any model obtained from either procedure. Notice however, the disparity in precision and recall – precision is much lower than recall. This is not surprising; inside-outside is known to converge to incorrect solutions for PCFGs (Lari and Young, 1990; de Marcken, 1995). This causes the *f*-score to deteriorate in successive iterations.

[5]This baseline is slightly lower than that reported in Table 1 due to holding out an additional 7000 sentences from the treebank training set. In order to accommodate unknown words from the test data (sec 23), the treebank model t_0 is smoothed in a manner similar to that shown in Eq. 9, with the test words (tagged using Treetagger) forming $g(w,\tau)$ and $\lambda = 0.1$. A testset is always merged with a given model in this manner before parsing, to account for unknown words.

		t_{0t}	It 1	It 2	It 3	It 4	It 5	It 6
Interleaved	Recall	86.48	86.72	***86.79**	*86.79	*86.78	86.81	86.72
Procedure	Precision	86.61	86.95	***87.07**	*87.06	*87.07	87.04	87.01
	f-score	86.55	86.83	***86.93**	*86.92	*86.92	86.92	86.86
Standard	Recall	86.48	87.95	87.11	86.42	85.55		
Procedure	Precision	86.61	85.99	84.79	83.37	82.06		
	f-score	86.5	86.96	85.93	84.87	83.77		

Table 2: Labeled Bracketing scores for various models, on PTB section 23.

The improvement in labeled bracketing *f*-score for the interleaved procedure is small, but is an encouraging result. The benefit to the re-estimated models comes only from better estimates of lexical parameters. We expect that re-estimation will benefit parameters associated with low frequency words - lexical parameters for high frequency words are bound to be estimated accurately from the treebank. We did not expect a large impact on labeled bracketing scores, given that low frequency words have correspondingly few occurrences in this test dataset. It is possible that the impact on *f*-score will be higher for a test set from a different domain. Note also that the size of our unlabeled training corpus (∼4M words) is relatively small – only about 4 times the PTB.

5 Verbal Subcategorization

We focus on learning verbal subcategorization, as a typical case of lexico-syntactic information. The subcategorization frame (SF) of verbs is a parameter of our PCFG - verbal tags in the PCFG are followed by an incorporation sequence that denotes the SF for that verb. We evaluate the re-estimated models on the task of detecting correct SFs of verbs in maximum-probability (viterbi) parses obtained using the models. All tokens of verbs and their preterminal symbols (consisting of a PoS tag and an incorporation sequence encoding the SF) are extracted from the viterbi parses of sentences in a testset. This tag-SF sequence is compared to a gold standard, and is scored correct if the two match exactly. PoS errors are scored as incorrect, even if the SF is correct. The gold standard is obtained from the transformed PTB trees.

The incorporation sequence corresponding to the SF consists of 3 features: The first one denotes basic categories of subcategorization such as transitive, intransitive, ditransitive, NP-PP, S, etc. The second feature denotes, for clausal complements, the type of clause (finite, infinite, small clause, etc.). The third feature encodes the nature of the subject of the clausal complements (empty category or non-empty). For example, the verb *considered* in the treebank sentence *They are officially considered strategic* gets a preterminal sequence of VBD.s.e.sc. This sequence indicates a past tense verb (VBD) with a clausal complement (s) which has an empty subject (e) since the sentence is passive and is of the type *small clause* (sc). A control verb (with an infinitival complement) in the sentence fragment *..did not want to fund X..* gets the frame s.e.to (see Fig. 1 for an example of a verb with its complement, as parsed by our PCFG). We have a total of 81 categories of SFs (without counting specific prepositions for prepositional frames), making fairly fine-grained distinctions of verbal categories.

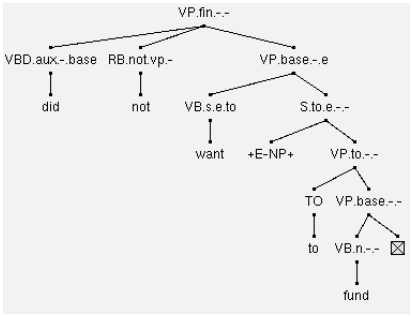

Figure 1: A subcat. frame for control verb *want*.

5.1 Learning Subcat Frames of Novel Verbs

We measure the error rate in the detection of the subcategorization frame of 1360 tokens of 117 verbs in Testset I. Recall from §3.2 that these verbs are novel verbs with respect to the treebank model. Table 3 shows this error rate (i.e. the fraction of test items which receive incorrect tag-incorporations in viterbi parses) for various models obtained using the interleaved and standard re-estimation procedures. t_{0t1} is the treebank model t_0 with the test data from Testset I merged in (to

Iteration i	Interleaved Procedure	Standard Procedure
t_{0t1}	33.36	33.36
1	*24.40	28.69
2	*23.45	25.56
3	*23.05	27.86
4	*22.89	28.41
5	***22.81**	-
6	*22.83	-

Table 3: Subcat. error for novel verbs (Testset I).

account for unknown words) using the smoothing scheme given in Eq. 9. This model has no verb specific information for the test verbs. For each test verb, it has a smoothed SF distribution proportional to the SF distribution for all verbs of that tag. The baseline error is 33.36%. This means that there is enough information in the average distribution of all verbs to correctly assign the subcategorization frame to novel verbs in 66.64% cases. For the models obtained using the interleaved re-estimation, the error rate falls to the lowest value of 22.81% for the model obtained in the 5^{th} iteration: an absolute reduction of 10.55 points, and a percentage error-reduction of 31.6%. The error reduction is statistically significant for all iterations compared to the baseline, with the 5^{th} iteration being also significantly better than the 1^{st}. The models obtained using standard re-estimation do not perform as well. Even for the model from the first iteration, whose labeled bracketing score was highest, the SF error is higher than the corresponding model from the interleaved procedure (possibly due to the low precision of this model). The error rate for the standard procedure starts to increase after the 2^{nd} iteration in contrast to the interleaved procedure.

5.2 Analysis of subcategorization learning

While the re-estimation clearly results in gains in SF detection for novel verbs, we also perform an evaluation for all verbs (novel and non-novel) in a given testset (Testset II as described in §3.2). The overall error reduction using the interleaved procedure is 8.97% (in Iteration 1). In order to better understand the relative efficacy of the supervised and unsupervised estimation for lexical items of different frequencies, we break up the set of test verbs into subsets based on their frequency of occurrence in the PTB training data, and evaluate them separately. Table 4 shows the error rates for verbs divided into these sets. We present error rates only for Iteration 1 in Table 4, since most of the error reduction takes place with the 1^{st} iteration. Statistically significant reductions are marked with * (confidence>99.9) and ** (>95). The second row shows error rates for verbs which have zero frequency in the treebank training data (i.e. novel verbs): Note that this error reduction is much less than the 31.6% in Testset I. These verbs are truly rare and hence have much fewer occurrences in the unlabeled corpus than Testset I verbs, which were artificially made novel (but are really mid-frequency verbs). This might indicate that error rates will decrease further if the size of the unlabeled corpus is increased. There is substantial error reduction for low-frequency verbs (<21 PTB occurrences). This is not hard to understand: the PTB does not provide enough data to have good parameter estimates for these verbs. For mid-to-high frequency verbs (from 21 to 500), the benefit of the unsupervised procedure reduces, though error reduction is still positive. Surprisingly, the error reduction for very high frequency verbs (more than 500 occurrences in the treebank) is also fairly high: we expected that parameters for high frequency words would benefit the least from the unsupervised estimation, given that they are already common enough in the PTB to be accurately estimated from it. The high frequency verbs (>500

TB Freq	t_{0t2}	It 1	Abs.Reduc	%Reduc
all	18.5	16.84	1.66	*8.97
0	41.26	33.01	8.25	*19.99
1	32.69	24.52	8.17	*24.99
2	36.55	22.76	13.79	*37.73
3	26.59	19.08	7.51	*28.24
4	22.38	20.28	2.1	9.38
5	24.63	19.40	5.23	*21.23
6-10	22.24	19.59	2.65	**11.92
11-20	21.54	18.02	3.52	*16.34
21-50	19.41	19.11	0.3	1.55
51-100	19.44	19.09	0.35	1.80
101-200	18.71	18.57	0.14	0.75
201-500	23.06	22.31	0.75	3.25
501-1K	18.07	16.82	1.25	6.92
1K-2K	12.38	12.25	0.13	1.05
2K-5K	9.42	7.62	1.8	*19.11
>5K	10.54	10.13	0.41	3.89

Table 4: Subcat. error breakup (Testset II)

occurrences) consist of very few types– mainly auxiliaries, some light verbs (*make*, *do*) and a few others (*rose*, *say*). It is possible that re-estimation from large data is beneficial for light verbs since they have a larger number of frames. The frequency range $2K$-$5K$ consists solely of auxiliary verbs. Examination of viterbi parses shows that improved results are largely due to better detection of predicative frames in re-estimated models.

To measure the impact of more unlabeled training data, we ran the interleaved procedure with 8M words of WSJ text. The SF error for novel verbs reduces to 22.06% in the 2^{nd} iteration (significantly different from the best error of 22.81% in the 5^{th} iteration for 4M words of training data)). We also get an improved overall error reduction of 9.9% on Testset II for the larger training data, as compared to 8.97% previously.

5.3 Previous Work

While there has been substantial previous work on the task of SF acquisition from corpora (Brent (1991); Manning (1993); Briscoe and Carroll (1997); Korhonen (2002), amongst others), we find that relatively few parsing-based evaluations are reported. Since their goal is to build probabilistic SF dictionaries, these systems are evaluated either against existing dictionaries, or on distributional similarity measures. Most are evaluated on testsets of *high*-frequency verbs (unlike the present work), in order to gauge the effectiveness of the acquisition strategy. Briscoe and Carroll (1997) report a token-based evaluation for seven verb types– their system gets an average recall accuracy of 80.9% for these verbs (which appear to be high-frequency verbs). This is slightly lower than the present system, which has an overall accuracy of 83.16% (on Testset II (It 1), as shown in Table 4). However, for low frequency verbs (exemplars $<$10) they report that their results are around chance. A parsing evaluation of their lexicon using an unlexicalized grammar as baseline, on 250 sentences from the Suzanne treebank gave a small (but not statistically significant) improvement in *f*-score (from 71.49 to 72.14%). Korhonen (2002) reports a parsing-based evaluation on 500 test sentences. She found a small increase in *f*-score (of grammatical relations markup) from 76.03 to 76.76. In general PARSEVAL measures are not very sensitive to subcategorization (Carroll et al., 1998); they therefore use a dependency-based evaluation. In the present research as well, we obtain statistically significant but quite small improvements in *f*-score (§4). Since we are interested in acquisition of PCFG lexicons, we focus our evaluations on verbal subcategorization of token occurrences of verbs in viterbi parses.

6 Conclusions

We have presented a methodology for incorporating additional lexical information from unlabeled data into an unlexicalized treebank PCFG. We obtain a large error reduction (31.6%) in SF detection for novel verbs as compared to a treebank baseline. The interleaved re-estimation scheme gives a significant increase in labeled bracketing scores from a relatively small unlabeled corpus. The interleaved scheme has an advantage over standard inside-outside PCFG estimation, as measured both by labeled bracketing scores and on the task of detecting SFs of novel verbs. Since our re-estimated models are treebank models, all evaluations are against treebank standards.

The grammar we worked with has very few incorporated features compared to the grammar used by, say Klein and Manning (2003). It would make sense to experiment with grammars with much richer sets of incorporated features. Features related to structure-selection by categories other than verbs – nouns, adverbs and adjectives – might be beneficial. These features should be incorporated as PCFG parameters, similar to verbal subcategorization. Experiments with 8 million words of training data gave significantly better results than with 4 million words, indicating that larger training sets will be beneficial as well. It would also be useful to make the transformation T of lexical parameters sensitive to treebank frequency of words. For instance, more weight should be given to the treebank model rather than the corpus model for mid-to-high frequency words, by making the parameter λ in T sensitive to frequency.

This methodology is relevant to the task of *domain-adaption*. Hara et al. (2007) find that retraining a model of HPSG lexical entry assignments is more critical for domain adaptation than re-training a structural model alone. Our PCFG captures many of the important dependencies captured in a framework like HPSG; in addition, we can use unlabeled data from a new domain in an unsupervised fashion for re-estimating lexical parameters, an important consideration in domain-adaption. Preliminary experiments on this task us-

ing New York Times unlabeled data with the PTB-trained PCFG show promising results.

Acknowledgments

I am grateful to Mats Rooth for extensive comments and guidance during the course of this research. The inside-outside re-estimation was conducted using the resources of the Cornell University Center for Advanced Computing.

References

F. Beil, G. Carroll, D. Prescher, S. Riezler, and M. Rooth. 1999. Inside-outside estimation of a lexicalized PCFG for German. In *Proceedings of the 37th meeting of ACL*.

Dan Bikel. 2004. Intricacies of Collins' Parser. *Computational Linguistics*, 30(4):479–511.

M. Brent. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th meeting of ACL*.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied NLP*.

G. Carroll and M. Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of EMNLP 1998*.

J. Carroll, G. Minnen, and E. Briscoe. 1998. Can subcategorization probabilities help parsing. In *Proceedings of 6th ACL/SIGDAT Workshop on Very Large Corpora*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of 43rd meeting of ACL*.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th meeting of ACL*.

Carl de Marcken. 1995. On the unsupervised induction of Phrase Structure grammars. In *3rd Workshop on Very Large Corpora*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Statistical Society*, 39(B):1–38.

Tejaswini Deoskar and Mats Rooth. 2008. Induction of Treebank-Aligned Lexical Resources. In *Proceedings of 6th LREC*.

Daniel Gildea. 2001. Corpus Variation and Parser Performance. In *Proceedings of EMNLP 2001*.

T. Hara, Y. Miyao, and J. Tsujii. 2007. Evaluating Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. In *Proceedings of the 10th International Conference on Parsing Technologies*.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 18(2):103–120.

Sabine Schulte imWalde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of LREC 2002*.

Mark Johnson. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4).

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*.

Anna Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, Univ. of Cambridge.

K. Lari and S. J. Young. 1990. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 4:35–56.

C. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st meeting of ACL*.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

D. McCloskey, E. Charniak, and M. Johnson. 2006. Effective Self-Training for Parsing. In *Proceedings of HLT-NAACL 2006*.

Pereira and Schabes. 1992. Inside-Outside reestimation from partially bracketed corpora. In *Proceedings of the 30th meeting of ACL*.

Detlef Prescher. 2003. A Tutorial on the Expectation-Maximization Algorithm Including Maximum-Likelihood Estimation and EM Training of Probabilistic Context-Free Grammars. ESSLLI 2003.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th COLING*.

Helmut Schmid. 2006. Trace Prediction and Recovery with Unlexicalised PCFGs and Slash Features. In *Proceedings of the 21st Conference on Computational Linguistics (COLING)*.