Smoothing fine-grained PCFG lexicons

Tejaswini Deoskar
ILLC
University of Amsterdam
t.deoskar@uva.nl

Mats Rooth
Dept. of Linguistics and CIS
Cornell University
mr249@cornell.edu

Khalil Sima'an
ILLC
University of Amsterdam
k.simaan@uva.nl

Abstract

We present an approach for smoothing treebank-PCFG lexicons by interpolating treebank lexical parameter estimates with estimates obtained from unannotated data via the Inside-outside algorithm. The PCFG has complex lexical categories, making relative-frequency estimates from a treebank very sparse. This kind of smoothing for complex lexical categories results in improved parsing performance, with a particular advantage in identifying obligatory arguments subcategorized by verbs unseen in the treebank.

1 Introduction

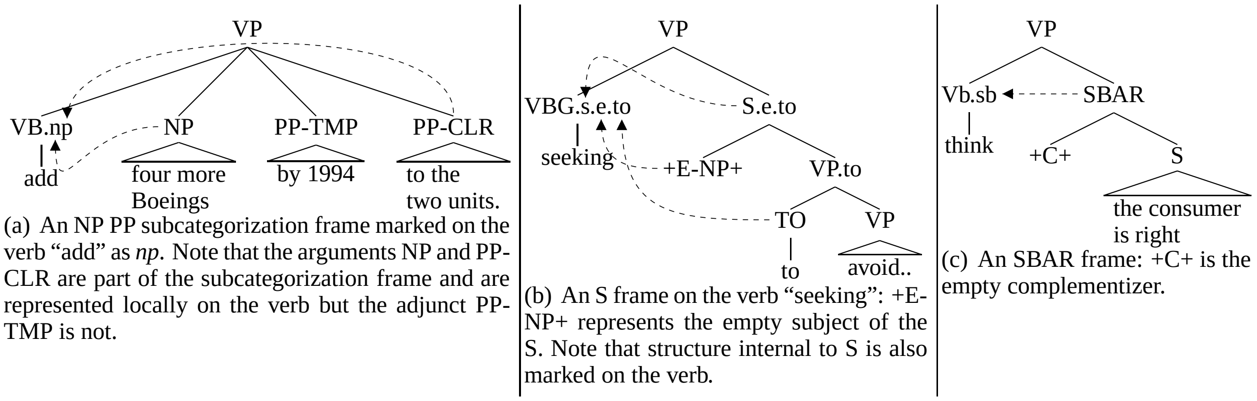
Lexical scarcity is a problem faced by all statistical NLP applications that depend on annotated training data, including parsing. One way of alleviating this problem is to supplement supervised models with lexical information from unlabeled data. In this paper, we present an approach for smoothing the lexicon of a treebank PCFG with frequencies estimated from unannotated data with Inside-outside estimation (Lari and Young, 1990). The PCFG is an unlexicalised PCFG, but contains complex lexical categories (akin to *supertags* in LTAG (Bangalore and Joshi, 1999) or CCG (Clark and Curran, 2004)) encoding structural preferences of words, like subcategorization.

The idea behind unlexicalised parsing is that the syntax and lexicon of a language are largely independent, being mediated by "selectional" properties of open-class words. This is the intuition behind lexicalised formalisms like CCG: here lexical categories are fine-grained and syntactic in nature. Once a word is assigned a lexical category, the word itself is not taken into consideration further in the syntactic analysis. Fine-grained categories imply that lexicons estimated from treebanks will be extremely sparse, even for a language like English with a large treebank resource like the Penn Treebank (PTB) (Marcus et al., 1993). Smoothing a treebank lexicon with an external wide-coverage lexicon is problematic due to their respective representations being incompatible and without an obvious mapping, assuming that the external lexicon is probabilistic to begin with. In this paper, we start with a treebank PCFG with fine-grained lexical categories and *re-estimate* its parameters on a large corpus of unlabeled data. We then use reestimates of lexical parameters (i.e. pre-terminal to terminal rule probabilities) to smooth the original treebank lexical parameters by interpolation between the two. Since the treebank PCFG itself is used to propose analyses of new data, the mapping problem is inherently taken care of. The smoothing procedure takes into account the fact that unsupervised estimation has benefits for unseen or low-frequency lexical items, but the treebank relative-frequency estimates are more reliable in the case of high-frequency items.

2 Treebank PCFG

In order to have fine-grained and linguistic lexical categories (like CCG) within a simple formalism with well-understood estimation methods, we first build a PCFG containing such categories from the PTB. The PCFG is unlexicalised (with limited lexicalization of certain function words, like in Klein and Manning (2003)). It is created by first transforming the PTB (Johnson, 1998) in an appropriate way and then extracting a PCFG from the transformed trees (Deoskar and Rooth, 2008). All functional tags in the PTB (such as NP-SBJ, PP-TMP, etc.) are maintained, as are all empty categories, making long-distance dependencies recoverable. The PCFG is trained on the standard training sections of the PTB and performs at the state-of-the-art level for unlexicalised PCFGs, giving 86.6% f-score on Sec. 23.

Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 214–217,
Paris, October 2009. ©2009 Association for Computational Linguistics



(a) An NP PP subcategorization frame marked on the verb "add" as *np*. Note that the arguments NP and PP-CLR are part of the subcategorization frame and are represented locally on the verb but the adjunct PP-TMP is not.

(b) An S frame on the verb "seeking": +E-NP+ represents the empty subject of the S. Note that structure internal to S is also marked on the verb.

(c) An SBAR frame: +C+ is the empty complementizer.

Figure 1: Subcategorized structures are marked as features on the verbal POS category.

An important feature of our PCFG is that pre-terminal categories for open-class items like verbs, nouns and adverbs are more complex than PTB POS tags. They encode information about the structure selected by the lexical item, in effect, its subcategorization frame. A pre-terminal in our PCFG consists of the standard PTB POS tag, followed by a sequence of features incorporated into it. Thus, each PTB POS tag can be considered to be divided into multiple finer-grained "supertags" by the incorporated features. These features encode the structure selected by the words. We focus on verbs in this paper, as they are important structural determiners. A sequence of one or more features forms the "subcategorization frame" of a verb: three examples are shown in Figure 1. The features are determined by a fully automated process based on PTB tree structure and node labels. There are 81 distinct subcategorization frames for verbal categories. The process can be repeated for other languages with a treebank annotated in the PTB style which marks arguments like the PTB.

3 Unsupervised Re-estimation

Inside-outside (henceforth I-O) (Lari and Young, 1990), an instance of EM, is an iterative estimation method for PCFGs that, given an initial model and a corpus of unannotated data, produces models that assign increasingly higher likelihood to the corpus at each iteration. I-O often leads to sub-optimal grammars, being subject to the well-known problem of local maxima, and dependence on initial conditions (de Marcken, 1995) (although there have been positive results using I-O as well, for e.g. Beil et al. (1999)). More recently, Deoskar (2008) re-estimated an unlexicalised PTB PCFG using unlabeled Wall Street Journal data. They compared models for which all PCFG parameters were re-estimated from raw data to models for which only lexical parameters were re-estimated, and found that the latter had better parsing results. While it is common to constrain EM either by good initial conditions or by heuristic constraints, their approach used syntactic parameters from a treebank model to constrain re-estimation of lexical parameters. Syntactic parameters are relatively well-estimated from a treebank, not being as sparse as lexical parameters. At each iteration, the re-estimated lexicon was interpolated with a treebank lexicon, ensuring that re-estimated lexicons did not drift away from the treebank lexicon.

We follow their methodology of constrained EM re-estimation. Using the PCFG with fine lexical categories (as described in §2) as the initial model, we re-estimate its parameters from an unannotated corpus. The lexical parameters of the re-estimated PCFG form its probabilistic "lexicon", containing the same fine-grained categories as the original treebank PCFG. We use this re-estimated "lexicon" to smooth the lexical probabilities in the treebank PCFG.

4 Smoothing based on a POS tagger : the initial model.

In order to use the treebank PCFG as an initial model for unsupervised estimation, new words from the unannotated training corpus must be included in it – if not, parameter values for new words will never be induced. Since the treebank model contains no information regarding correct feature sequences for unseen words, we assign all possible sequences that have occurred in the treebank model with the POS tag of the word. We assign all possible sequences to *seen* words as well – although the word is seen, the correct feature sequence for a structure in a training sentence might still be unseen with that word. This is done as follows: a standard POS-tagger (TreeTagger, (Schmid, 1994)) is used to tag the unlabeled corpus. A frequency table $c_{pos}(w, \tau)$ consisting of words and POS-tags is extracted from the resulting corpus, where w is the word and τ its POS tag. The frequency $c_{pos}(w, \tau)$ is split amongst all possible feature sequences ι for that POS tag in proportion to treebank marginals $t(\tau, \iota)$ and $t(\tau)$

$$c_{pos}(w, \tau, \iota) = \frac{t(\tau, \iota)}{t(\tau)} c_{pos}(w, \tau) \tag{1}$$

Then the treebank frequency $t(w, \tau, \iota)$ and the scaled corpus frequency are interpolated to get a smoothed model t_{pos}. We use λ=0.001, giving a small weight initially to the unlabeled corpus.

$$t_{pos}(w, \tau, \iota) = (1 - \lambda) t(w, \tau, \iota) + \lambda c_{pos}(w, \tau, \iota) \tag{2}$$

The first term will be zero for words unseen in the treebank: their distribution in the smoothed model will be the average treebank distribution over all possible feature sequences for a POS tag. For seen words, the treebank distribution over feature sequence is largely maintained, but a small frequency is assigned to unseen sequences.

5 Smoothing based on EM re-estimation

After each iteration i of I-O, the expected counts $c_{em_i}(w, \tau, \iota)$ under the model instance at iteration $(i - 1)$ are obtained. A smoothed treebank lexicon t_{em_i} is obtained by linearly interpolating the smoothed treebank lexicon $t_{pos}(w, \tau, \iota)$ and a scaled re-estimated lexicon $\bar{c}_{em_i}(w, \tau, \iota)$.

$$t_{em_i}(w, \tau, \iota) = (1-\lambda) t_{pos}(w, \tau, \iota) + \lambda \bar{c}_{em_i}(w, \tau, \iota) \tag{3}$$

where $0 < \lambda < 1$. The term $\bar{c}_{em_i}(w, \tau, \iota)$ is obtained by scaling the frequencies $c_{em_i}(w, \tau, \iota)$ obtained by I-O, ensuring that the treebank lexicon is not swamped with the large training corpus[1].

$$\bar{c}_{em_i}(w, \tau, \iota) = \frac{t(\tau, \iota)}{\sum_w c_{em_i}(w, \tau, \iota)} c_{em_i}(w, \tau, \iota) \tag{4}$$

λ determines the relative weights given to the treebank and re-estimated model for a word. Since parameters of high-frequency words are likely to be more accurate in the treebank model, we parametrize λ as λ_f according to the treebank frequency $f = t(w, \tau)$.

6 Experiments

The treebank PCFG is trained on sections 0-22 of the PTB, with 5000 sentences held-out for evaluation. We conducted unsupervised estimation using Bitpar (Schmid, 2004) with unannotated Wall Street Journal data of 4, 8 and 12 million words, with sentence length <25 words. The treebank and re-estimated models are interpolated with $\lambda = 0.5$ (in Eq. 3). We also parametrize λ for treebank frequency of words – optimizing over a development set gives us the following values of λ_f for different ranges of treebank word frequencies.

$$\begin{array}{ll} \text{if } t(w,\tau) <= 5 \,, & \lambda_f = 0.5 \\ \text{if } 5 < t(w,\tau) <= 15 \,, & \lambda_f = 0.25 \\ \text{if } 15 < t(w,\tau) <= 50 \,, & \lambda_f = 0.05 \\ \text{if } t(w,\tau) > 50 \,, & \lambda_f = 0.005 \end{array} \tag{5}$$

Evaluations are on held-out data from the PTB by stripping all PTB annotation and obtaining Viterbi parses with the parser Bitpar. In addition to standard PARSEVAL measures, we also evaluate parses by another measure specific to subcategorization[2]: the POS-tag+feature sequence on verbs in the Viterbi parse is compared against the corresponding tag+feature sequence on the transformed PTB gold tree, and errors are counted. The tag-feature sequence correlates to the structure selected by the verb, as exemplified in Fig. 1.

7 Results

There is a statistically significant improvement[3] in labeled bracketing f-score on Sec. 23 when the treebank lexicon is smoothed with an EM-re-estimated lexicon. In Table 1, t_t refers to the baseline treebank model, smoothed using the POS-tag smoothing method (from §4) on the test data (Sec. 23) in order to incorporate new words from the test data[4]. t_{pos} refers to the initial model for re-estimation, obtained by smoothed the treebank model with the POS-tag smoothing method with the large unannotated corpus (4 million words). This model understandably does not improve over t_t for parsing Sec. 23. $t_{em_1, \lambda=0.5}$ is the model obtained by smoothing with an EM-re-estimated model with a constant interpolation factor $\lambda = 0.5$. This model gives a statistically significant improvement in f-score over both t_t and t_{pos}. The last model t_{em_1, λ_f} is obtained by smoothing with

[1] Note that in Eq. 4, the ratio of the two terms involving c_{em_i} is the conditional, lexical probability $P_{em_i}(w|\tau, \iota)$.

[2] PARSEVAL measures are known to be insensitive to subcategorization (Carroll et al., 1998).

[3] A randomized version of a paired-sample t-test is used.

[4] This is always done before parsing test data.

	t_t	t_{pos}	$t_{em_1,\lambda=0.5}$	t_{em_1,λ_f}
Recall	86.48	86.48	86.72	87.44
Precision	86.61	86.63	86.95	87.15
f-score	86.55	86.56	*86.83	*87.29

Table 1: Labeled bracketing F-score on section 23.

an interpolation factor as in Eq. 5 : this is the best model with a statistically significant improvement in f-score over t_t, t_{pos} and $t_{em_1,\lambda=0.5}$.

Since we expect that smoothing will be advantageous for unseen or low-frequency words, we perform an evaluation targeted at identifying structures subcategorized by unseen verbs. Table 2 shows the error reduction in identifying subcat. frames in Viterbi parses, of unseen verbs and also of all verbs (seen and unseen) in the testset. A breakup of error by frame type for unseen verbs is also shown (here, only frames with >10 token occurrences in the *test* data are shown). In all cases (unseen verbs and all verbs) we see a substantial error reduction. The error reduction improves with larger amounts of unannotated training data.

8 Discussion and Conclusions

We have shown that lexicons re-estimated with I-O can be used to smooth unlexicalised treebank PCFGs, with a significant increase in f-score even in the case of English with a large treebank resource. We expect this method to have more impact for languages with a smaller treebank or richer tag-set. An interesting aspect is the substantial reduction in subcategorization error for unseen verbs for which no word-specific information about subcategorization exists in the unsmoothed or POS-tag-smoothed lexicon. The error reduction in identifying subcat. frames implies that some constituents (such as PPs) are not only attached correctly but also identified correctly as arguments (such as PP-CLR) rather than as adjuncts.

There have been previous attempts to use POS-tagging technologies (such as HMM or maximum-entropy based taggers) to enhance treebank-trained grammars (Goldberg et al. (2009) for Hebrew, (Clark and Curran, 2004) for CCG). The re-estimation method we use builds full parse-trees, rather than use local features like taggers do, and hence might have a benefit over such methods. An interesting option would be to train a "supertagger" on fine-grained tags from the PTB and to supertag a large corpus to harvest lexical frequencies. This would form another (possibly higher) baseline for the I-O re-estimation approach presented here and is the focus of our future work.

Frame	# tokens	%Error	%Error	%Error
	(test)	t_{pos}	t_{em_1}	Reduc.
All unseen (4M words)	1258	33.47	22.81	31.84
All unseen (8M words)	1258	33.47	22.26	33.49
All unseen (12M words)	1258	33.47	21.86	34.68
transitive	662	23.87	18.73	21.52
intransitive	115	38.26	33.91	11.36
NP PP-CLR	121	34.71	32.23	7.14
PP-CLR	73	27.4	20.55	25
SBAR	124	12.1	12.1	0
S	12	83.33	58.33	30
NP NP	10	90	80	11.11
PRT NP	21	38.1	33.33	12.5
s.e.to (see Fig.1b)	50	16	12	25
NP PP-DIR	11	63.64	54.55	14.28
All verbs (4M)	11710	18.5	16.84	8.97

Table 2: Subcat. error for verbs in Viterbi parses.

References

S. Bangalore and A. K. Joshi. 1999. Supertagging: An Approach to Almost Parsing. *Computational Linguistics*, 25:237–265.

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

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

S. Clark and J. R. Curran. 2004. The Importance of Supertagging for Wide-Coverage CCG Parsing. In *22nd COLING*.

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

T. Deoskar. 2008. Re-estimation of Lexical Parameters for Treebank PCFGs. In *22nd COLING*.

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

Y. Goldberg, R. Tsarfaty, M. Adler, and M. Elhadad. 2009. Enhancing Unlexicalized Parsing Performance using a Wide Coverage Lexicon, Fuzzy Tag-set Mapping, and EM-HMM-based Lexical Probabilities. In *EACL-09*.

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

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *ACL 41*.

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.

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.

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

H. Schmid. 2004. Efficient Parsing of Highly Ambiguous CFGs with Bit Vectors. In *20th COLING*.