Active Zipfian Sampling for Statistical Parser Training*

Onur Çobanoğlu
Department of Computer Science
Sennott Square
University of Pittsburgh
Pittsburgh, PA 15260, USA
onc3@pitt.edu

Abstract

Active learning has proven to be a successful strategy in quick development of corpora to be used in training of statistical natural language parsers. A vast majority of studies in this field has focused on estimating informativeness of samples; however, representativeness of samples is another important criterion to be considered in active learning. We present a novel metric for estimating representativeness of sentences, based on a modification of Zipf's *Principle of Least Effort*. Experiments on WSJ corpus with a wide-coverage parser show that our method performs always at least as good as and generally significantly better than alternative representativeness-based methods.

1 Introduction

Wide coverage statistical parsers (Collins, 1997; Charniak, 2000) have proven to require large amounts of manually annotated data for training to achieve substantial performance. However, building such large annotated corpora is very expensive in terms of human effort, time and cost (Marcus et al., 1993). Several alternatives of the standard supervised learning setting have been proposed to reduce the annotation costs, one of which is active learning. Active learning setting allows the learner to select its own samples to be labeled and added to the training data iteratively. The motive behind active learning is that if the learner may select highly informative samples, it can eliminate the redundancy generally found in random data; however, informative samples can be very untypical (Tang et al., 2002). Unlike random sampling, active learning has no guarantee of selecting *representative* samples and untypical training samples are expected to degrade test performance of a classifier.

To get around this problem, several methods of estimating representativeness of a sample have been introduced. In this study, we propose a novel representativeness estimator for a sentence, which is based on a modification of Zipf's *Principle of Least Effort* (Zipf, 1949), theoretically sound and empirically validated on Brown corpus (Francis and Kučera, 1967). Experiments conducted with a wide coverage CCG parser (Clark and Curran, 2004; Clark and Curran, 2007) on CCGbank (Hockenmaier and Steedman, 2005) show that using our estimator as a representativeness metric never performs worse than and generally outperforms length balanced sampling (Becker and Osborne, 2005), which is another representativeness based active learning method, and pure informativeness based active learning.

2 Related Work

In selective sampling setting, there are three criteria to be considered while choosing a sample to add to the training data (Dan, 2004; Tang et al., 2002): *Informativeness* (what will the expected contribution of this sample to the current model be?), *representativeness* (what is the estimated probability of seeing this sample in the target population?) and *diver-*

*Vast majority of this work was done while the author was a graduate student in Middle East Technical University, under the funding from TÜBİTAK-BİDEB through 2210 National Scholarship Programme for MSc Students.

Proceedings of NAACL HLT 2009: Short Papers, pages 249–252,
Boulder, Colorado, June 2009.

sity (how different are the samples in a batch from each other?). The last criterion applies only to the batch-mode setting, in which the training data is incremented by multiple samples at each step for practical purposes.

Most of the active learning research in statistical parser training domain has focused on informativeness measures developed for both single and multi-learner settings. The informativeness measures for single-learners that have exhibited significant performance in well known experimental domains are as follow: Selecting the sentences unparsable by the current model (and if the batch does not get filled, using a secondary method) (Thompson et al., 1999); selecting the sentences with the highest *tree entropy*, i.e. the Shannon entropy of parses the probabilistic parser assigns to the sentence (Hwa, 2004); selecting the sentences having *lowest best probabilities*, where *best probability* is the conditional probability of the most probable parse, given the sentence and the current model (Osborne and Baldridge, 2004); primarily selecting the sentences that are expected to include events observed with low frequency so far with the help of bagging and filling the rest of the batch according to tree entropy, which is named as *two-stage active learning* by Becker and Osborne (2005). Proposed informativeness measures for multiple learners and *ensemble* learners can be found in (Baldridge and Osborne, 2003; Osborne and Baldridge, 2004; Becker and Osborne, 2005; Baldridge and Osborne, 2008).

As for representativeness measures, Tang et. al. (2002) proposed using *sample density*, i.e. the inverse of the average distance of the sample to the other samples in the pool, according to some distance metric. Becker and Osborne (2005) introduced *length balanced sampling*, in which the length histogram of the batch is kept equal to the length histogram of a random sample of batch size drawn from the pool.

3 Description Of The Work

We introduce a novel representativeness measure for statistical parser training domain. Our measure is a function proposed in (Sigurd et al., 2004), which estimates the relative frequencies of sentence lengths in a natural language. Sigurd et. al. (2004) claimed that the longer a sentence is, the less likely it will be uttered; in accordance with Zipf's Principle of Least Effort (Zipf, 1935). However, too short sentences will appear infrequently as well, since the number of different statements that may be expressed using relatively fewer words is relatively smaller. Authors conjectured that there is a clash of expressivity and effort over the frequency of sentence length, which effort eventually wins. They formulated this behavior with a Gamma distribution estimating the relative frequencies of sentence lengths. Authors conducted a parameter fit study for English using Brown corpus (Francis and Kučera, 1967) and reported that the formula $f(L) = 1.1 \times L^1 \times 0.90^L$, where L is the sentence length, fits to the observations with very high correlation.

We propose using this fitted formula (named $f_{zipf-eng}$ from now on) as the measure of representativeness of a sentence. This metric has several nice features: It is model-independent (so it is not affected from modeling errors), is both theoretically sound and empirically validated, can be used in other NLP domains and is a numerical metric, providing flexibility in combining it with informativeness (and diversity) measures.

4 Experiments

4.1 Experimental Setup

We conducted experiments on CCGbank corpus (Hockenmaier and Steedman, 2005) with the wide coverage CCG parser of Clark and Curran (2004; 2007)[1]. C&C parser was fast enough to enable us to use the whole available training data pool for sample selection in experiments, but not for training (since training C&C parser is not that fast). Among the models implemented in the parser, the normal-form model is used. We used the default settings of the C&C parser distribution for fair evaluation. WSJ Sections 02-21 (39604 sentences) are used for training and WSJ Section 23 (2407 sentences) is used for testing. Following (Clark and Curran, 2007), we evaluated the parser performance using the labeled f-score of the predicate-argument dependencies produced by the parser.

[1]Following (Baldridge and Osborne, 2004), we claim that the performances of AL with C&C parser and other state-of-the-art wide coverage parsers will be similar

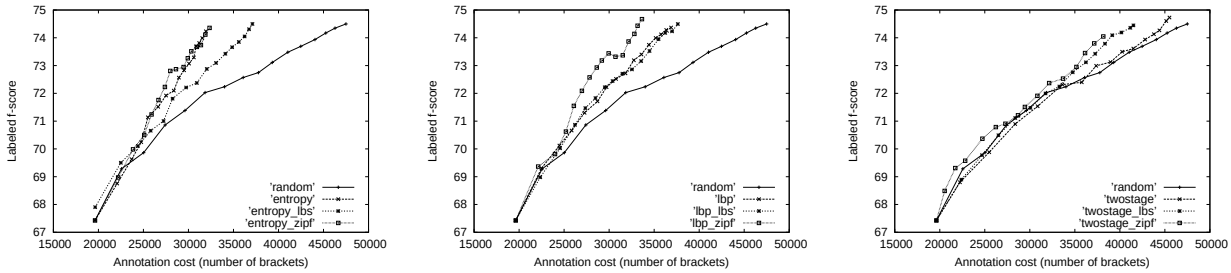

Figure 1: Comparative performances of different representativeness measures. The informativeness measure used is tree entropy in the leftmost graph, lowest best probability in the central graph and two-stage AL in the rightmost graph. The line with the tag 'random' always shows the random sampling baseline.

	none	**lbs**	**zipf**	**random**
entropy	30.99%(74.24%)	20.63%(74.31%)	30.11%(74.36%)	N/A (74.35%)
lbp	22.34%(74.37%)	20.78%(74.49%)	30.19%(74.43%)	N/A(74.50%)
unparsed/entropy	19.98%(74.32%)	19.34%(74.43%)	26.27%(74.38%)	N/A(74.35%)
twostage	2.83%(73.94%)	11.13%(74.09%)	13.38%(74.05%)	N/A(73.94%)

Table 1: PRUD values of different AL schemes. The row includes the informativeness measure and the column includes the representativeness measure used. The column with the label **random** always includes the results for random sampling. The numbers in parentheses are the labeled f-score values reached by the schemes.

For each active learning scheme and random sampling, the size of the seed training set is 500 sentences, the batch size is 100 sentences and iteration stops after reaching 2000 sentences.[2] For statistical significance, each experiment is replicated 5 times. We evaluate the active learning performance in terms of *Percentage Reduction in Utilized Data*, i.e. how many percents less data is used by AL compared to random sampling, in order to reach a certain performance score. Amount of used data is measured with the number of brackets in the data. In CCGbank, a bracket always corresponds to a parse decision, so it is a reasonable approximation of the amount of annotator effort.

Our measure is compared to length balanced sampling and using no representativeness measures. Since there is not a trivial distance metric between CCG parses and we do not know a proposed one, we could not test it against sample density method. We limited the informativeness measures to be tested to the four single-learner measures we mentioned in Section 2. Multi-learner and ensemble methods are excluded, since the success of such methods relies heavily on the diversity of the available models (Baldridge and Osborne, 2004; Baldridge and Osborne, 2008). The models in C&C parser are not diverse enough and we left crafting such diverse models to future work.

We combined $f_{zipf-eng}$ with the informativeness measures as follow: With tree entropy, sentences with the highest $f_{zipf-eng}(s) \times f_{nte}(s, G)$ (named $f_{zipf-entropy}(s, G)$) values are selected. $f_{nte}(s, G)$ is the tree entropy of the sentence s under the current model G, normalized by the binary logarithm of the number of parses, following (Hwa, 2004). With lowest best probability, sentences with the highest $f_{zipf-eng}(s) \times (1 - f_{bp}(s, G))$ values are selected, where f_{bp} is the best probability function (see Section 2). With unparsed/entropy, we primarily chose the unparsable sentences having highest $f_{zipf-eng}(s)$ values and filled the rest of the batch according to $f_{zipf-entropy}$. With two-stage active learning, we primarily chose sentences that can be parsed by the full parser but not the bagged parser and have the highest $f_{zipf-eng}(s)$ values, we secondarily chose sentences that cannot be parsed by both parsers and have the highest $f_{zipf-eng}(s)$ values, the third priority is given to sentences having highest

[2]These values apply to the training of the parser and the CCG supertagger. POS-tagger is trained with the whole available pool of 39604 sentences due to sparse data problem.

$f_{zipf-entropy}$ values.[3] Combining length balanced sampling with all of these informativeness measures is straightforward. For statistical significance, a different random sample is used for length histogram in each replication of experiment.

4.2 Results

Results can be seen in Figure 1 and Table 1. Due to lack of space and similarity of the graphs of unparsed/entropy and LBP, we excluded the graph of unparsed/entropy (but its results are included in Table 1). Since observation points in different lines do not fall on the exactly same performance level (for exact PRUD measurement), we took the points on as closest f-score levels as possible. With tree entropy, Zipfian sampling performs almost as good as pure informativeness based AL and with two-stage AL, length balanced sampling performs almost as good as Zipfian sampling. In all other comparisons, Zipfian sampling outperforms its alternatives substantially.

5 Conclusion and Future Work

We introduced a representativeness measure for active learning in statistical parser training domain, based on an empirical sentence length frequency model of English. Experiments on a wide coverage CCG parser show that this measure outperforms the alternative measures most of the time and never hinders. Our study can be extended via further experimentation with the methods we excluded in Section 4.1, with other parsers, with other languages and with other Zipfian cues of language (e.g. Zipf's law on word frequencies (Zipf, 1949)).

Acknowledgments

We specially thank to Jason Baldridge, Cem Bozşahin, Ruken Çakıcı, Rebecca Hwa, Miles Osborne and anonymous reviewers for their invaluable support, advices and feedback.

References

Jason Baldridge and Miles Osborne. 2003. Active learning for HPSG parse selection. In *Proceedings of CoNLL*.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of EMNLP*.

Jason Baldridge and Miles Osborne. 2008. Active learning and logarithmic opinion pools for HPSG parse selection. In *Natural Language Engineering*, volume 14, pages 199–222. Cambridge, UK.

Markus Becker and Miles Osborne. 2005. A two-stage method for active learning of statistical grammars. In *Proceedings of IJCAI*.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of ACL*.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

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

Shen Dan. 2004. Multi-criteria based active learning for named entity recognition. Master's thesis, National University of Singapore.

W. Nelson Francis and Henry Kučera. 1967. *Computational Analysis of Present-day American English*. Brown University Press, Providence, RI.

Julia Hockenmaier and Mark Steedman. 2005. *CCGbank*. Linguistic Data Consortium, Philadelphia.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3):253–276.

Mitchell P. Marcus, Mary A. Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English:The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miles Osborne and Jason Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proceedings of HLT-NAACL*.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost van Weijer. 2004. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, 58(1):37–52.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of ACL*.

Cynthia A. Thompson, Mary E. Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of ICML*.

George K. Zipf. 1935. *The Psychobiology of Language*. MIT Press, Cambridge, MA. Reprinted in 1965.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

[3] Note that our usage of two-stage AL is slightly different from the original definition in (Becker and Osborne, 2005)