Memory-based active learning for French broadcast news

Frédéric Tantini, Christophe Cerisara, Claire Gardent

LORIA-CNRS
Campus Scientifique, 54506 Vandoeuvre-les-Nancy
{tantini,cerisara,gardent}@loria.fr

Abstract

Stochastic dependency parsers can achieve very good results when they are trained on large corpora that have been manually annotated. Active learning is a procedure that aims at reducing this annotation cost by selecting as few sentences as possible that will produce the best possible parser. We propose a new selective sampling function for Active Learning that exploits two memory-based distances to find a good compromise between parser uncertainty and sentence representativeness. The reduced dependency between both parsing and selection models opens interesting perspectives for future models combination. The approach is validated on a French broadcast news corpus creation task dedicated to dependency parsing. It outperforms the baseline uncertainty entropy-based selective sampling on this task. We plan to extend this work with self- and co-training methods in order to enlarge this corpus and produce the first French broadcast news Tree Bank.

1. Introduction

Syntactic parsing is a key component of most natural language processing applications, which commonly exploit nowadays stochastic dependency parsers trained on large Tree Banks. Despite the importance of dependency parsing, no such corpora exist to the best of our knowledge for French broadcast news parsing. Most efforts on French parsing are based on the French Tree Bank (FTB) [1] that contains newspaper texts. Yet, it is well-known that cross-domain parsing is a difficult challenge, and our preliminary experiments show that porting a parser trained on the FTB to a broadcast news corpus dramatically reduces the parsing accuracy from 88% to 55%. Our main objective is thus to build a new Tree Bank on top of the broadcast news ESTER corpus [2], which we call Ester Tree Bank (ETB) [3]. The first steps towards this objective involves annotating a small initial bootstrapping corpus, and enlarging this corpus with semi-supervised approaches. We investigate in this work the use of active learning for this purpose, and propose a memory-based selective sampling method for dependency parsing that combines sample knowledge with density.

Our baseline corpus is the ESTER corpus [2], which contains manual transcriptions of French broadcast news utterances. This corpus is originally designed for speech recognition evaluation, and the transcription guidelines are adapted to match the capabilities of speech recognizers. In particular, disfluencies are annotated as follows:

- Every acoustic realization that corresponds to a lexicon entry is transcribed. This includes repetitions and hesitations ("uh... ").
- Conversely, false starts and incomplete words are not transcribed
- For the same reason, punctuation is not transcribed, and every word is in lower-case, except for acronyms and proper nouns.

In this work, we use for part-of-speech tagging the TreeTagger software [4] and for parsing the state-of-the-art Malt parser [5].

Semi-supervised training algorithms exploit both a small annotated and a large unannotated corpus to train a classification or parsing model. Active Learning is such an iterative training approach that chooses at each iteration a few examples to annotate manually, in order to maximize the performances of the resulting parsing model, hence minimizing the work of annotators. The most famous Active Learning approaches are first reviewed in section 2. Then, an original selection criterion is proposed in section 3. The proposed approach is evaluated in section 4 and section 5 concludes the paper.

2. Review of Active Learning for syntactic parsing

We focus in this section on pool-based active learning, also known as selective sampling, which selects the next example to manually annotate from some unlabeled corpus [6]. Although the literature about active learning for classification is huge, it is much smaller for the application of active learning to parsing. We thus briefly review next in priority the active learning works dedicated to parsing, although we may also cite other active learning works when discussing general active learning concepts.

One of the most common Active Learning frameworks is uncertainty sampling [7], which selects examples for which the classifier is the least confident, where confidence is typically derived from the entropy or other margin metrics [8]. Various simple criteria (sentence length, infrequent words, . . .) are also studied in [9], amongst them sentence length has been found to be the best one. Conversely, it is shown in [10] that sentence-length-based selection criteria performs poorly for the application of Active Learning to statistical parsers. The author has also compared uncertainty and likelihood-based selection and concluded in favor of the former. Uncertainty is there classically represented by the entropy computed on the set of possible parses returned by the parser. Furthermore, a "lexical novelty" measure that computes the number of unseen co-occurring word-pairs is also proposed but is not developed further on because of its weak performances. We propose in this paper to investigate a related but more complex lexical measure.

An alternative to these approaches is to combine several models, for instance in the framework of co-learning. Hence, Dredze *et al.* [11] introduce confidence estimation in margin-based Active Learning approaches, and Tang *et al.* [12] propose to first cluster unlabeled data using kmeans, and then query the

most uncertain sentences of each cluster.

Finally, some works study the Active Learning protocol itself. In particular, the missed-cluster effect is an undesirable side-effect of Active Learning [13]: when a given cluster has no initial examples, this cluster is far from the boundaries of the learning and its examples are thus considered as reliable, when they are not. This effect is largely reduced when annotating complete sentences instead of just single words, as unseen clusters' examples might co-occur with known class boundaries.

The baseline Active Learning method that we have used in this work is uncertainty sampling based on the entropy of the class posterior distribution, which is a common choice in many related works [14, 15, 16]. However, a classical issue in uncertainty sampling is the selection of outlier examples. This issue is addressed in [17] and [18] by optimizing an estimation of the classification error instead of uncertainty, in [15] by combining uncertainty with prototypicality or in [14] through the proposed *sampling by uncertainty and density (SUD)* paradigm. We follow the same approach than the latter work on SUD, except that we compute the density over the whole unlabeled corpus, that we propose a memory-based distance instead of a more traditional uncertainty measure, and that we apply this combined criterion on dependency structures parsing.

Please refer to [19] and [20] for further recent surveys on Active Learning.

3. Active Learning approaches

We compare in this section four selection criteria: the baseline random selection, an approximation of the upper-bound oracle selection, the baseline uncertainty-based selective sampling and a memory-based selection. We show that the best results are obtained with the proposed memory-based selection system.

3.1. Training procedure

Our implementation of the Active Learning training procedure exploits three independent data sets, respectively for learning (L), development (U) and testing (T). L contains all training instances that have been labeled manually, U only contains unlabeled instances and T is the gold standard used for evaluation. At each iteration, a new parsing model is trained on L and is evaluated on T. Then, a single unlabeled sentence is chosen from U based on a given selection criterion, is manually annotated, and moved into L: the pseudo-code for this process is given in Alg. 1.

Algorithm 1: Pseudo-code

Input: Two sets L and U of labeled and unlabeled examples, a testing set T, a *select* function

repeat
- `// Learn a model using` L
- $\lambda \leftarrow$ train(L);
- test(λ, T);
- `// Query some unlabeled data`
- $x \leftarrow$*select*(U);
- `// Annotate and move the data`
- $U \leftarrow U \setminus \{x\}$;
- $L \leftarrow L \cup \{x\}$;

until *some stopping criterion*;

Active Learning is especially useful for very small values of L, i.e. at the very beginning of any corpus creation process, because it is an efficient approach for building a first set of models that reach a minimum level of accuracy. This nicely fits our corpus bootstrapping objective, and our experimental conditions match these expectations, as shown in section 4. Once good enough models are available, increasing the size of the corpus shall be realized with different approaches, such as bootstrapping, self- and co-training, corpus mixing, etc.

3.2. Approximated oracle selection

The optimal selection function in Alg.1 is the one for which the learning curve has the steepest ascent, i.e., the highest derivative value at the initial iterations first. As proposed in [17], the first-order Markov assumption allows to approximate this oracle ordering by a deterministic process that iteratively selects the single unlabeled sentence that maximizes parsing performances:

$$\text{select}_{oracle}(U) = \arg\max_{s \in U} \left(\text{Score}_{\lambda(L \cup s)}(T)\right)$$

where $\lambda(L \cup s)$ represents the parameters of the parsing model trained on $L \cup s$, and $\text{Score}_{\lambda(L \cup s)}(T)$ is the Labeled Attachment Score (*LAS*), as defined in the CoNLL evaluations, which measures the ratio of words with correct predicted governor and dependency type on the test set T.

This approximated oracle learning curve gives the upper limit, or best reachable performances, of all the selection criteria evaluated next. Conversely, we also build the baseline learning curve by randomly choosing the next sentence:

$$\text{select}_{baseline}(U) = \text{Random}(s \in U)$$

3.3. Uncertainty-based sampling

Our uncertainty baseline is the log-loss approach proposed in [17], which presents the advantage of combining both statistical and pragmatic active learning strategies [18] and is commonly used as a standard baseline for uncertainty-based sampling. With this approach, the error is estimated by the entropy of the class posterior distribution. Although this entropy can be easily computed for a Bayesian classifier, it is much more difficult to estimate for a stochastic parsing process, which manipulates structured data. In this case, an approach proposed in [10] consists in estimating the entropy from the set of all possible parses. However, this is not possible in our case, because our parser only produces a single parse for each sentence.

Our chosen parser is the Malt parser, which is a state-of-the-art stochastic dependency parser [5] that incrementally applies a sequence of actions on two word stacks in order to build the final dependency structure d_s on the sequence of words $s = (w_1, \cdots, w_N)$. Initially, the words (w_i) are all pushed into the first stack, and subsequent *actions* (a_i) manipulate these stacks by shifting the top word from the first to the second stack, deleting the top word of the second stack, or adding a dependency relation between both top words. The process terminates when the first stack is empty. This stochastic process transforms the final tree posterior into the probability of a sequence of *actions*:

$$\begin{aligned} P(d_s|s, \lambda(L)) &= P(a_1, \cdots, a_T|s, \lambda(L)) \\ &\simeq \prod_{i=1}^{T} P(a_i|s, \lambda(L)) \end{aligned} \quad (1)$$

when assuming independence of the actions. In the Malt parser, a Support Vector Machine (SVM) classifier is trained to associate a set of features, which encode the current parsing state, to

the best action that leads to the reference parse. SVMs do not directly return class posterior probabilities, but several methods exist to estimate such probabilities in the multi-class case. We have chosen the approach described in [21] and implemented in Libsvm [22]. Then, the entropy of the posterior distribution is computed for each individual posterior $P(a_i|s, \lambda(L))$ and averaged over all (a_i) in the parse:

$$H(s) = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j} P(a_i = j|s, \lambda(L)) \log\left(P(a_i = j|s, \lambda(L))\right)$$

The sentence chosen by this uncertainty-based baseline is the one that maximizes this entropy:

$$\text{select}_{uncertainty}(U) = \arg\max_{s \in U} H(s)$$

3.4. Memory-based sentence selection

We propose next a new selection criterion that looks for a compromise between model uncertainty and representativeness of the next sentence to label.

As discussed in section 2, this involves selecting the next sentence to label based on two criteria: the uncertainty about its estimated parse - we want to choose sentences that are not already well-modeled, and the representativeness of this sentence - we want to avoid selecting outliers that are not representative of the rest of the corpus.

Uncertainty is classically estimated from the trained model itself (see section 3.3). However, such a self-estimation of classification confidence suffers from the same limitations and bias than the model itself: a totally erroneous posterior distribution might lead to a high confidence in its result. We argue that confidence measures may benefit from being estimated using radically different classification models than the one that has been trained in the first place. We thus propose to estimate the classification uncertainty of our SVM model using a memory-based distance. Our hypothesis is that the parser will tend to be more confident for utterances that are close to the ones in the training corpus, while the resulting parse is more likely to be wrong when the test utterance is very different from every known training sentence. The model is thus the least certain for the utterance s that maximizes this distance to the training corpus:

$$\hat{s} = \max_{s \in U}\left(\min_{s' \in T} d(s, s')\right)$$

where $d(s, s')$ is the Levenshtein distance between the sequence of part-of-speech (POS) tags of respectively s and s'.

Another related approach might be to further compare the estimated parse trees, in addition to POS-tags. However, such a distance would increase the dependency between both models, which may increase the correlation between their respective uncertainty estimation errors.

Representativeness is modeled in previous work by data density in a neighborhood of s in U [14, 12, 23]. We rather propose to measure the representativeness of s by its average Levenshtein distance to the whole corpus U. This global measure shall thus select in priority the sentences that have the largest impact on the average performances of the parser. The combined criterion is finally:

$$\text{select}_{mbl}(U) = \arg\max_{s \in U}\left(\min_{s' \in T} d(s, s') - \left(\frac{1}{|U|}\sum_{s' \in U} d(s, s')\right)\right)$$

The relative importance of both distances has not been tuned in this linear combination.

4. Experimental validation

The corpus used is the Ester corpus [2], which contains 37 hours of manually transcribed French broadcast news. The following experiments are based on a small part of this corpus, composed of 19591 words manually annotated with syntactic dependencies. This corpus is randomly split into three sets: a learning set L (5% of sentences), a developing set U (85%), and a testing set T (10%). Evaluations are realized with 10-fold cross-validation, leading to a confidence interval of $\pm 0.6\%$.

We compare next the four systems defined previously: oracle and random (section 3.2), uncertainty (section 3.3) and MBL (section 3.4). The chosen evaluation metric is the Labeled Attachment Score (*LAS*). It is a standard measure from the CoNLL evaluations [24]. Fig. 1 shows the learning curve, i.e., the Labeled Attachment Score as a function of the number of words added in the training corpus.

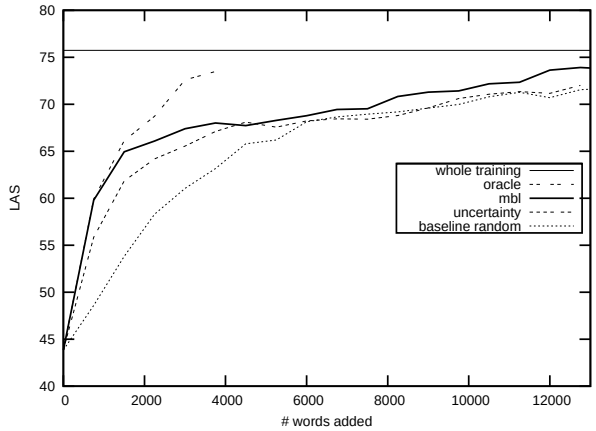

Figure 1: Comparison of the training curves for the four selection approaches.

The top horizontal line represents the performances obtained when training on the whole corpus. Because of its very high computational complexity, we have not been able to realize as much iterations with the oracle system than with the others, which explains the relatively shorter curve. Yet, the oracle curve gives a good idea of the potential of selective sampling, and we can also observe that the proposed approach gives comparable performances than the oracle system at the very initial stage of Active Learning, up to a training size of 1500 words, which is a very good result. The proposed system is also the best one for all iterations.

The *deficiency* measure has been proposed to quantify Active Learning performances (see for instance [25]) by integrating and computing the ratio of the areas above the learning curves: the smaller it is, the better is the system; 1 corresponds to similar areas between the proposed algorithm and the random baseline. It is here of 0.53 for the MBL selective sampling and 0.73 for uncertainty sampling: this compares relatively well to the gains reported in [25].

The main Active Learning objective is to reach the best performances with as few training examples as possible. The proposed approach is indeed especially interesting on the first third of our corpora, and we expect this behavior to scale up with the

size of the unlabeled corpus considered. However, once reasonably good accuracy is reached, we plan to consider alternative approaches to Active Learning, such as self- and co-training, in order to further speed up enlarging the ETB corpus. Also, the edit distance used so far is based on part-of-speech tags, which may be adequate with a small initial corpus size, but which might also tend towards zero with an increasing size of the training corpus, leading to a reduced discriminative power between correct and incorrect sentences. It might thus be better for larger corpora to compute edit distances based on more precise POS-tags, lemmas or inflected forms.

5. Conclusions

We have proposed in this work a new memory-based selective sampling criterion for Active Learning of a stochastic dependency parser. The proposed selection function combines an uncertainty with a representativeness measures, in order to circumvent the classical issue of outliers selection in uncertainty-based sampling. Contrary to classical Active Learning approaches, the uncertainty measure is estimated from a totally different model than the parser, which reduces the issue of self-estimation of the confidence of a classifier. It further facilitates future models combination for uncertainty estimation. The proposed representativeness measure is a global measure on the unlabeled corpus derived from the Levenshtein distance between part-of-speech sequences. This Active Learning algorithm has been compared favorably with the classical uncertainty baseline computed from the entropy of the class posterior distribution, averaged over all local decisions taken by the stochastic parser. We have not considered so far the human cost required to annotate sentences, even though such a cost is often taken into account in Active Learning. But on initial conditions such as the ones used here, annotation costs do not play a critical role, and we have thus preferred to focus first on the decrease of parsing errors.

This work is the first one in the process of building a French tree bank dedicated to broadcast news. The next steps shall involve enlarging this bootstrapping corpus with self- or co-training approaches, with the objective of reaching 150 000 words and 84% of LAS score at a low cost.

6. References

[1] M.-H. Candito, B. Crabb, P. Denis, and F. Gurin, "Analyse syntaxique du français : des constituants aux dé pendances," in *Actes de TALN*, Senlis, 2009.

[2] G. Gravier, J.-F. Bonastre, S. Galliano, douard Geoffrois, K. M. Tait, and K. Choukri, "Ester, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques," in *Proceedings Journes d'tude sur la Parole*, Fez, 2004.

[3] C. Cerisara and C. Gardent, "Analyse syntaxique du français parlé," in *Journée thmatique ATALA : Quels analyseurs syntaxiques pour le français ?*, 2009.

[4] H. Schmid, "Improvements in part-of-speech tagging with an application to german," in *Proc. of the ACL SIGDAT-Workshop*, Dublin, 1995, pp. 47–50.

[5] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi, "Maltparser: A language-independent system for data-driven dependency parsing," *Natural Language Engineering*, vol. 13, no. 2, pp. 95–135, 2007, http://w3.msi.vxu.se/ nivre/research/MaltParser.html.

[6] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon University, May 2005.

[7] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 3–12.

[8] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: an evaluation," *Mach. Learn.*, vol. 68, no. 3, pp. 235–265, 2007.

[9] K. Ohtake, "Analysis of selective strategies to build a dependency-analyzed corpus," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 635–642.

[10] R. Hwa, "Sample selection for statistical parsing," *Comput. Linguist.*, vol. 30, no. 3, pp. 253–276, 2004.

[11] M. Dredze and K. Crammer, "Active learning with confidence," in *HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. Morristown, NJ, USA: Association for Computational Linguistics, 2008, pp. 233–236.

[12] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," in *ACL*, 2002, pp. 120–127.

[13] K. Tomanek, F. Laws, U. Hahn, and H. Schütze, "On proper unit selection in active learning: Co-selection effects for named entity recognition," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 9–17. [Online]. Available: http://www.aclweb.org/anthology/W09-1902

[14] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," 2008, pp. 1137–1144.

[15] W. Daelemans, H. Groenewald, and G. Van Huyssteen, "Prototype-based active learning for lemmatization," in *Proceedings of Recent Advances in Natural Language Processing*, 2009, pp. 65–70.

[16] G. Druck, B. Settles, and A. McCallum, "Active learning by labeling features," 2009, pp. 81–90.

[17] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th International Conf. on MachineLearning*, 2001, pp. 441–448.

[18] M. Li and I. K. Sethi, "Confidence-based active learning," *IEEE trans. on pattern analysis and machine intelligence*, vol. 28, no. 8, pp. 1251–1261, Aug. 2006.

[19] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[20] "Hlt '09: Proceedings of the naacl hlt 2009 workshop on active learning for natural language processing." Morristown, NJ, USA: Association for Computational Linguistics, 2009, conference Chair-Ringger, Eric and Conference Chair-Hertel, Robbie and Conference Chair-Tomanek, Katrin.

[21] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Maching Learning Research*, vol. 5, pp. 975–1005, 2004.

[22] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[23] D. Shen, J. Zhang, J. Su, G. Zhou, and C.-L. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proc. 42nd annual meeting on Association for Computational Linguistics*, 2004.

[24] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre, "The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies," in *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, Manchester, United Kingdom, 2008, pp. 159–177.

[25] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2008, pp. 1137–1144.