A Web Survey on the Use of Active Learning to Support Annotation of Text Data

Katrin Tomanek
Jena University Language & Information Engineering Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, D-07743 Jena, Germany
katrin.tomanek@uni-jena.de

Fredrik Olsson
SICS
Box 1263
SE-164 29 Kista, Sweden
fredrik.olsson@sics.se

Abstract

As supervised machine learning methods for addressing tasks in natural language processing (NLP) prove increasingly viable, the focus of attention is naturally shifted towards the creation of training data. The manual annotation of corpora is a tedious and time consuming process. To obtain high-quality annotated data constitutes a bottleneck in machine learning for NLP today. Active learning is one way of easing the burden of annotation. This paper presents a first probe into the NLP research community concerning the nature of the annotation projects undertaken in general, and the use of active learning as annotation support in particular.

1 Introduction

Supervised machine learning methods have been successfully applied to many NLP tasks in the last few decades. While these techniques have shown to work well, they require large amounts of labeled training data in order to achieve high performance. Creating such training data is a tedious, time consuming and error prone process. Active learning (AL) is a supervised learning technique that can be used to reduce the annotation effort. The main idea in AL is to put the machine learner in control of the data from which it learns; the learner can ask an oracle (typically a human) about the labels of the examples for which the model learned so far makes unreliable predictions. The active learning process takes as input a set of labeled examples, as well as a larger set of unlabeled examples, and produces a classifier and a relatively small set of newly labeled data. The overall goal is to create as good a classifier as possible, without having to mark-up and supply the learner with more data than necessary. AL aims at keeping the human annotation effort to a minimum, only asking the oracle for advice where the training utility of the result of such a query is high. Settles (2009) gives a detailed overview of the literature on AL.

It has been experimentally shown that AL can indeed be successfully applied to a range of NLP tasks including, e.g., text categorization (Lewis and Gale, 1994), part-of-speech tagging (Dagan and Engelson, 1995; Ringger et al., 2007), parsing (Becker and Osborne, 2005), and named entity recognition (Shen et al., 2004; Tomanek et al., 2007). Despite that somewhat impressive results in terms of reduced annotation effort have been achieved by such studies, it seems that AL is rarely applied in real-life annotation endeavors.

This paper presents the results from a web survey we arranged to analyze the extent to which AL has been used to support the annotation of textual data in the context of NLP, as well as addressing the reasons to why or why not AL has been found applicable to a specific task. Section 2 describes the survey in general, Section 3 introduces the questions and presents the answers received. Finally, the answers received are discussed in Section 4.

2 The Survey

The survey was realized in the form of a web-based questionnaire; the primary reason for this approach, as opposed to reading and compiling information

Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, pages 45–48,
Boulder, Colorado, June 2009.

from academic publications, was that we wanted to free ourselves and the participants from the dos and don’ts common to the discourse of scientific papers.

The survey targeted participants who were involved in the annotation of textual data intended for machine learning for all kinds of NLP tasks. It was announced on the following mailing lists: BioNLP, Corpora, UAI List, ML-news, SIG-IRlist, Linguist list, as well as lists reaching members of SIGANN, SIGNLL, and ELRA. By utilizing these mailing lists, we expect to have reached a fairly large portion of the researchers likely to participate in annotation projects for NLP. The questionnaire was open February 6–23, 2009.

After an introductory description and one initial question, the questionnaire was divided into two branches. The first branch was answered by those who had used AL to support their annotation, while the second branch was answered by those who had not. Both branches shared a common first part about the general set-up of the annotation project under scrutiny. The second part of the AL-branch focused on experiences made with applied AL. The second part of the non AL-branch asked questions about the reasons why AL had not been used. Finally, the questionnaire was concluded by a series of questions targeting the background of the participant.

The complete survey can be downloaded from `http://www.julielab.de/ALSurvey`.

3 Questions and answers

147 people participated in the survey. 54 completed the survey while 93 did not, thus the overall completion rate was 37 %. Most of the people who did not complete the questionnaire answered the first couple of questions but did not continue. Their answers are not part of the discussion below. We refrain from a statistically analysis of the data but rather report on the distribution of the answers received.

Of the people that finished the survey, the majority (85 %) came from academia, with the rest uniformly split between governmental organizations and industry. The educational background of the participants were mainly computational linguistics (46 %), general linguistics (22 %), and computer science (22 %).

3.1 Questions common to both branches

Both the AL and the non-AL branch were asked several questions about the set-up of the annotation project under scrutiny. The questions concerned, e.g., whether AL had been used to support the annotation process, the NLP tasks addressed, the size of the project, the constitution of the corpus annotated, and how the decision when to stop the annotation process was made.

The use of AL as annotation support. The first question posed was whether people had used AL as support in their annotation projects. 11 participants (20 %) answered this question positively, while 43 (80 %) said that they had not used AL.

The task addressed. Most AL-based annotation projects concerned the tasks information extraction (IE) (52 %), document classification (17.6 %), and (word sense) disambiguation (17.6 %). Also in non AL-based projects, most participants had focused on IE tasks (36.8 %). Here, syntactic tasks including part-of-speech tagging, shallow, and deep parsing were also often considered (19.7 %). Textual phenomena, such as coreferences and discourse structure (9.6 %), and word sense disambiguation (5.5 %) formed two other answer groups. Overall, the non AL-based annotation projects covered a wider variety of NLP tasks than the AL-based ones. All AL-based annotation projects concerned English texts, whereas of the non-AL projects only 62.8 % did.

The size of the project. The participants were also asked for the size of the annotation project in terms of number of units annotated, number of annotators involved and person months per annotator. The average number of person months spent on non AL-projects was 21.2 and 8.7 for AL-projects. However, these numbers are subject to a high variance.

The constitution of the corpus. Further, the participants were asked how the corpus of unlabeled instances was selected.[1] The answer options included *(a)* taking all available instances, *(b)* a random subset of them, *(c)* a subset based on keywords/introspection, and *(d)* others. In the AL-branch, the answers were uniformly distributed be-

[1] The unlabeled instances are used as a pool in AL, and as a corpus in non AL-based annotation.

tween the alternatives. In the non AL-branch, the majority of participants had used alternatives *(a)* (39.5 %) and *(b)* (34.9 %).

The decision to stop the annotation process. A last question regarding general annotation project execution concerned the stopping of the annotation process. In AL-based projects, evaluation on a held-out gold standard (36.5 %) and the exhaustion of money or time (36.5 %) were the major stopping criteria. Specific stopping criteria based on AL-internal aspects were used only once, while in two cases the annotation was stopped because the expected gains in model performance fell below a given threshold.

In almost half (47.7 %) of the non AL-based projects the annotation was stopped since the available money or time had been used up. Another major stopping criterion was the fact that the complete corpus was annotated (36 %). Only in two cases annotation was stopped based on an evaluation of the model achievable from the corpus.

3.2 Questions specific to the AL-branch

The AL-specific branch of the questionnaire was concerned with two aspects: the learning algorithms involved, and the experiences of the participants regarding the use of AL as annotation support. Percentages presented below are all related to the 11 persons who answered this branch.

Learning algorithms used. As for the AL methods applied, there was no single most preferred approach. 27.3 % had used uncertainty sampling, 18.2 % query-by-committee, another 18.2% error reduction-based approaches, and 36.4 % had used an “uncanonical” or totally different approach which was not covered by any of these categories. As base learners, maximum-entropy based approaches as well as Support-Vector machines were most frequently used (36.4 % each).

Experiences. When asked about their experiences, the participants reported that their expectations with respect to AL had been partially (54.4 %) or fully (36.3 %) met, while one of the participants was disappointed. The AL participants did not leave many experience reports in the free text field. From the few received, it was evident that the sampling complexity and the resulting delay or idle time of the annotators, as well as the interface design are critical issues in the practical realization of AL as annotation support.

3.3 Question specific to the non-AL branch

The non AL-specific branch of the questionnaire was basically concerned with why people did not use AL as annotation support and whether this situation could be changed. The percentages given below are related to the 43 people who answered this particular part of the questionnaire.

Why was not AL used? Participants could give multiple answers to this question. Many participants had either never heard of AL (11 %) or did not use AL due to insufficient knowledge or expertise (26 %). The implementational overhead to develop an AL-enabled annotation editor kept 17.8 % of the participants from using AL. Another 19.2 % of the participants stated that their project specific requirements did not allow them to use AL. Given the comments given in the free text field, it can be deduced that this was often the case when people wanted to create a corpus that could be used for a multitude of purposes (such as building statistics on, cross-validation, learning about the annotation task per se, and so forth) and not just for classifier training. In such scenarios, the sampling bias introduced by AL is certainly disadvantageous. Finally, about 20.5 % of the participants were not convinced that AL would work well in their scenario or really reduce annotation effort. Some participants stated in their free form comments that while they believed AL would reduce the amount of instances to be annotated it would probably not reduce the overall annotation time.

Would you consider using AL in future projects? According to the answers of another question of the survey, 40 % would in general use AL, while 56 % were sceptical but stated that they would possibly use a technique such as AL.

4 Discussion

Although it cannot be claimed that the data collected in this survey is representative for the NLP research community as a whole, and the number of participants was too low to draw statistically firm conclusions, some interesting trends have indeed been discovered within the data itself. The conclusions drawn in this section are related to the answers provided in light of the questions posed in the survey.

The questionnaire was open to the public and was not explicitly controlled with respect to the distribution of characteristics of the sample of the community that partook in it. One effect of this, coupled with the fact that the questionnaire was biased towards those familiar with AL, is that we believe that the group of people that have used AL are overrepresented in the data at hand. However, this cannot be verified. Nevertheless, given this and the potential reach of the mailing lists used for announcing the survey, it is remarkable that not more than 20 % (11 out of 54) of the participants had used AL as annotation support.

The doubts of the participants who did not use AL towards considering the technique as a potential aid in annotation in essence boil down to the absence of an AL-based annotation editor, as well as the difficulty in estimating the effective reduction in effort (such as time, money, labor) that the use of AL imply. Put simply: Can AL for NLP really cut annotation costs? Can AL for NLP be practically realized without too much overhead in terms of implementation and education of the annotator? Research addressing the former question is ongoing which is shown, e.g., by the recent Workshop on Cost-Sensitive Learning held in conjunction with the Neural Information Processing Systems Conference 2008. As for the latter question, there is evidently a need of a general framework for AL in which (specialized) annotation editors can be realized. Also, hand-in-hand with the theoretical aspects of AL and their practical realizations in terms of available software packages, there clearly is a need for usage and user studies concerning the effort required by human annotators operating under AL-based data selection schemes in real annotation tasks.

Two things worth noticing among the answers from participants of the survey that had used AL include that most of these participants had positive experiences from using AL, although turn-around time and consequently the idle time of the annotator remains a critical issue; and that English was the only language addressed. This is somewhat surprising given that AL seems to be a technique well suited for bootstrapping language resources for, e.g., so called "under resourced" languages. Also we were surprised by the fact that both in AL and non-AL projects rather "unsophisticated" criteria were used to decide about the stopping of annotation projects.

Acknowledgements

The first author was funded by the German Ministry of Education and Research within the STEMNET project (01DS001A-C) and the EC within the BOOTStrep project (FP6-028099). We wish to thank Björn Gambäck for commenting and proofreading drafts of this paper.

References

Markus Becker and Miles Osborne. 2005. A two-stage method for active learning of statistical grammars. In *Proc. of the 19th International Joint Conference on Artificial Intelligence*, pages 991–996.

Ido Dagan and Sean P. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proc. of the 12th International Conference on Machine Learning*, pages 150–157.

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proc. of the Linguistic Annotation Workshop*, pages 101–108.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 589–596.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 486–495.