Efficient and Effective Text-Annotation through Active Learning

Markus Zlabinger
markus.zlabinger@tuwien.ac.at
Vienna University of Technology
Vienna, Austria

ABSTRACT

The most commonly used active learning criterion is uncertainty sampling, where a supervised model is used to predict uncertain samples that should be labeled next by a human annotator. When using active learning, two problems are encountered: First, the supervised model has a cold-start problem in the beginning of the active learning process, and second, the human annotators might make labeling mistakes. In my Ph.D. research, I address the two problems with the development of an unsupervised method for the computation of informative samples. The informative samples are first manually labeled and then used for both: The training of the initial active learning model and the training of the human annotators in form of a learning-by-doing session. The planned unsupervised method will be based on word-embeddings and it will be limited to the area of text classification.

CCS CONCEPTS

• **Computing methodologies** → **Active learning settings**; • **Applied computing** → **Annotation**.

KEYWORDS

text annotation; active learning; crowdsourcing

ACM Reference Format:
Markus Zlabinger. 2019. Efficient and Effective Text-Annotation through Active Learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3331184.3331424

1 ACTIVE LEARNING INITIALIZATION

The supervised model in an active learning (AL) process has a cold-start problem since no labeled samples are available right in the beginning of the process. To address this problem, several papers suggest methods for the unsupervised computation of samples (so-called *seed samples*) used for the initialization of the active learning model. The methods proposed in the area of text classification [2, 3] often consider word-based features (e.g. TF-IDF weighted word-vectors) for the computation of seed samples. The disadvantage of word-based features is that two words that are semantically similar, like *"purchase"* and *"buy"*, are considered dissimilar.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331424

2 METHOD

In the scope of my Ph.D. research, I intend to develop a new unsupervised method based on word-embeddings for the computation of *seed samples*. The method takes a set of unlabeled samples as input from which it computes the seed samples. Afterwards, the seed samples are manually labeled and used for the training of the initial active learning model.

The method will be based on two main concepts: First, *diversity* where the idea is that a model can learn more from two different samples than from two similar samples. And second, *polarization* where the idea is that a model can learn more from a sample that contains a high fraction of polarizing words (e.g. "successful", "tremendous"), than from a sample that is less polarizing.

I intend to experimentally compare the proposed method to the previous methods (e.g. [2, 3]) on various text classification datasets. My hypothesis is that seed samples computed with the proposed method are more effective for the initialization of an active learning model (i.e. higher initial prediction accuracy).

3 ANNOTATOR TRAINING

In the active learning process, a potential error source is the human annotator that assigns labels for uncertain samples [1]. To reduce the number of errors (and therefore increase the label-quality), I propose to train the human annotator based on a set of already labeled samples from which the annotator can learn how a given annotation task should be performed. I intend to answer two research questions: First, what samples should be used for the training of the human annotators (e.g. AL seed samples, a set of manually compiled samples)? And second: How are the training samples effectively presented to the annotators (e.g. learning-by-doing)? The focus will be on domains in which annotation is challenging (e.g. medical).

The experiments will be conducted with expert annotators and non-expert annotators (e.g. crowdsourcing), and the goal is to improve the label-quality of non-experts so that it approaches the quality of experts. Since the effect of annotator training is difficult to measure between individuals, I intend to measure the effect between groups of annotators (training vs. non-training group).

REFERENCES

[1] Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the Effects of Selective Sampling on the Annotation Task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning - CONLL '05.* Association for Computational Linguistics, Ann Arbor, Michigan, 144. https://doi.org/10.3115/1706543.1706569

[2] Rong Hu, Brian Mac Namee, and Sarah Jane Delany. 2010. Off to a Good Start: Using Clustering to Select the Initial Training Set in Active Learning.. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference.*

[3] Weiwei Yuan, Yongkoo Han, Donghai Guan, Sungyoung Lee, and Young-Koo Lee. 2011. Initial Training Data Selection for Active Learning. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication.* ACM, 5.