TEXTPRO-AL: An Active Learning Platform for Flexible and Efficient Production of Training Data for NLP Tasks

Bernardo Magnini[1], Anne-Lyse Minard[1,2], Mohammed R. H. Qwaider[1], Manuela Speranza[1]
[1] Fondazione Bruno Kessler, Trento, Italy
[2] Dept. of Information Engineering, University of Brescia, Italy
{magnini,minard,qwaider,manspera}@fbk.eu

Abstract

This paper presents TEXTPRO-AL (Active Learning for Text Processing), a platform where human annotators can efficiently work to produce high quality training data for new domains and new languages exploiting Active Learning methodologies. TEXTPRO-AL is a web-based application integrating four components: a machine learning based NLP pipeline, an annotation editor for task definition and text annotations, an incremental re-training procedure based on active learning selection from a large pool of unannotated data, and a graphical visualization of the learning status of the system.

1 Background and Motivations

Text Mining technologies are becoming more and more requested, as they work "behind the shoulder" of widespread applications: search engines adopt semantic strategies to match user needs, virtual assistants provide help in task-driven conversations, trends on social media are discovered and analyzed in huge amounts of data. These applications take advantage of the recent progresses in Computational Linguistics, which, to a large extent, are based on a massive use of Machine Learning (ML) technology for Natural Language Processing (NLP) tasks.

A key aspect motivating our proposal is that ML systems need training data (i.e. annotated corpora), which in turn are based on high quality manual linguistic annotations. As a matter of fact, manual production of datasets for training is still a core step for developing concrete NLP applications and, as a consequence, there is a high demand for methodologies that make the process more flexible and efficient.

Specifically, we are interested in the following issues: (i) applications require high flexibility in the use of different labeling categories (e.g. general categories like `Person` as opposed to fine-grained categories like `Football-Player`); (ii) in addition, domain adaptation requires that a dataset developed for a general domain (e.g. calendar dates for news) is reused for a more specific domain (e.g. the legal domain) without loosing performance; (iii) there is an increasing demand for applications supporting different languages, some of which might not be well covered in terms of annotated data; (iv) finally, it is current practice in research (particularly in shared evaluation tasks) to develop training data independently of the performance they allow to obtain in a certain task, although this is not optimal for the production cycle of applications. In concrete cases, training data are updated and revised incrementally till performance for the task at hand is satisfactory.

2 Active Learning

The key choice in designing TEXTPRO-AL was to make use of Active Learning (AL) (Cohn et al., 1994; Settles, 2010) as the core technology for optimizing training production. The main principle underlying AL is that the selection of the textual portions to be manually annotated is much more effective when it is guided by strong criteria (typically, informativeness, representativeness, and diversity of selected instances) than when it is performed randomly, as in standard supervised learning.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 131–135, Osaka, Japan, December 11-17 2016.

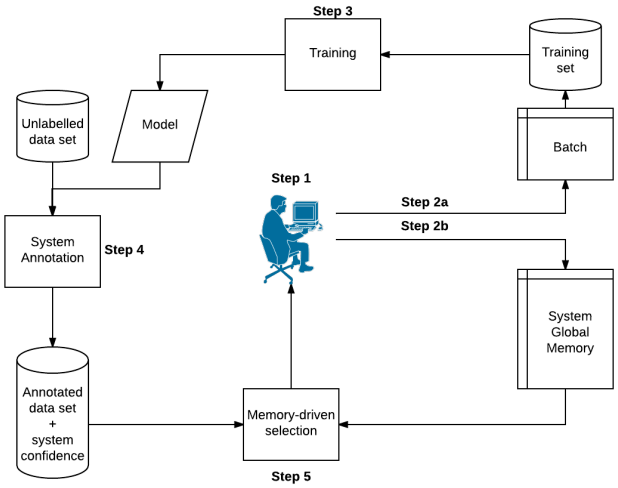

Figure 1: TEXTPRO-AL architecture.

These criteria are typically applied in an iterative way, following a re-training procedure, where instances are selected from a (usually large) pool of unlabeled texts. Although there are experimental evidences that AL allows for a significant reduction of the amount of training needed to achieve a certain performance (e.g. calculated in terms of F-measure), there is less experience and less consensus about the use of AL in practical contexts (Tomanek and Olsson, 2009).

In our implementation, the AL cycle (see Figure 1) starts with a human annotator providing supervision on a sample that has been tagged automatically by the system (step 1): the annotator is asked to either confirm the annotation (in case it is correct) or to revise it. The annotated instance is stored in a batch (step 2a), where it is accumulated with other instances for re-training and, as a result, a new model is produced (step 3). This model is used to automatically annotate a set of unlabeled documents and to assigns an estimated confidence score to each annotated instance (step 4).

In step 2b the manually supervised instance is stored in the Global Memory of the system (together with the revisions performed by the annotator). In step 5 a single instance is selected from the unlabeled dataset. The selected instance, as well as its relevant context, is removed from the unlabeled set and is presented to the human annotator to be revised.

Active Learning has been successfully experimented for a large variety of sequence labeling annotation tasks (an incomplete list includes (Shen et al., 2004) for Named Entity Recognition, (Ringger et al., 2007) for Part-of-Speech Tagging, and (Schohn and Cohn, 2000) for Text Classification), which guarantees the high portability of the approach we propose.

3 System Description

TEXTPRO-AL integrates four components in a single platform: (1) a ML-based *NLP pipeline*; (2) a web-based *annotation editor* for manually revising linguistic annotations; (3) an *AL package* which selects samples to be annotated from a large pool of non annotated documents and then re-trains the pipeline; (4) a *visualizer* of the internal learning status of the pipeline for the task at hand.

1. *NLP pipeline*. This is a set of tools for automatic text annotation based on ML classification. We assume that the pipeline is already available for a number of NLP tasks (e.g. part-of-speech tagging, named entity recognition), and that for each task the ML classifier already implements corresponding feature extractors (e.g. orthographic features for named entity recognition). We also assume that there are no hard coded linguistic categories for the NLP tasks (e.g. `Person` for NER), so that the pipeline builds a model for a labeling task by taking the categories directly from the training data. There are several linguistic pipelines of this type available, including CoreNLP[1] developed at

[1] `http://stanfordnlp.github.io/CoreNLP/`

Stanford University, the OpenNLP pipeline[2] and LingPipe[3]. For our demonstrator we use TextPro[4] (Pianta et al., 2008), a pipeline for English and Italian including several annotation layers, such as part-of-speech tagging, lemmatization, named entities recognition, dependency parsing and event extraction.

In order for a pipeline to be integrable with TEXTPRO-AL, it has to be able to produce an output in the IOB2 format[5] and to assign a confidence score to each labeled sequence.[6]

2. *Annotation editor*. This is a tool for manually inserting and revising linguistic annotations on a corpus. Required basic functions are the possibility to define a set of categories to be used for a certain annotation task and the capability to annotate a sequence of tokens with a certain category. Several open source annotation tools are available (e.g. Callisto,[7] WebAnno,[8] Brat,[9] and CAT[10]); among these, we selected MTEqual[11] (Girardi et al., 2014) (a tool developed for assessing the quality of machine translations) to integrate it in the current demonstrator, as it offers good editing features for online revisions. The use of MTEqual allows us to experiment the TEXTPRO-AL approach virtually on any sequence labeling annotation task.

3. *AL package*. This is a package for Active Learning which optimizes the selection of samples (from a large pool of unlabeled data) to be given for revision to the annotator. Only a small number of packages for AL are available (e.g. JCLAL[12]) and we preferred our own implementation, which is specifically targeted to NLP tasks.

4. *Learning visualizer*. This is a set of graphical tools allowing the annotator to monitor the learning status of the system. Specifically, we use learning curves produced with the Chart.js graphical package[13]. A learning curve shows the annotator the impact of the annotations on the performance of the system.

4 Novelty and Impact of the Platform

The TEXTPRO-AL platform aims at facilitating and making more efficient the development of training data for NLP tasks based on statistical machine learning. The goal is to give final users (e.g. companies) a platform which: (i) reduces the effort required to produce high quality training data; (ii) allows for easy and effective domain adaptation of existing classifiers; (iii) allows to monitor the performance of the classifier as the training data are incremented.

The technological novelty of the platform is the integration of three components, usually developed independently, in a single platform. To the best of our knowledge, this is the first system where a ML classifier, an annotation tool, and an active learning package are fully integrated. Particularly, while the role of AL has been scientifically investigated in controlled settings (e.g. (Shen et al., 2004) for named entity recognition, (Ringger et al., 2007) for part-of-speech tagging), the proposed platform allows for scientific experiments and uses in real settings, typically characterized by the presence of a huge (and uncontrolled) pool of unlabeled data.

Developing training data for new domains and new languages is of utmost importance for almost any text mining applications. As a consequence, reducing the time needed for data preparation may have

[2] http://opennlp.apache.org/index.html

[3] http://alias-i.com/lingpipe/index.html

[4] http://textpro.fbk.eu/

[5] The IOB2 tagging format is a common format for text chunking. B- is used to tag the beginning of a chunk, I- to tag tokens inside the chunk and O to indicate tokens not belonging to a chunk.

[6] Confidence scores can be obtained in terms of probabilities (e.g. with CRF algorithms), distance between a feature vector and the hyperplan (e.g. with SVM algorithms), etc.

[7] https://github.com/mitre/callisto

[8] https://webanno.github.io/webanno/

[9] http://brat.nlplab.org/index.html

[10] http://dh.fbk.eu/resources/cat-content-annotation-tool

[11] https://github.com/hltfbk/MT-EQuAl

[12] https://sourceforge.net/projects/jclal/

[13] http://www.chartjs.org

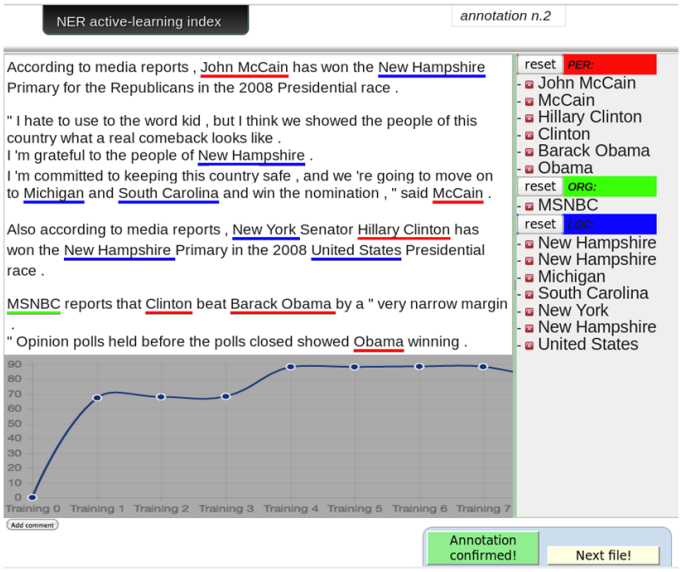

Figure 2: User interface of the TEXTPRO-AL platform.

a relevant impact on the overall production cycle of an application. In addition, the proposed platform enhances the user's experience by offering the human annotator the possibility of monitoring the impact of her/his work on the performance of the system.

5 The Platform at Work

The platform targets researchers and companies who need to develop training data for NLP applications, either from scratch or by extending existing datasets. The typical user of the system is a domain expert, whose goal is to produce a training set for a certain NLP task. In our experience, a user only needs a short training phase (in the order of one day) and some pilot annotations to learn how to use the system for most of the NLP tasks (e.g. part of speech tagging, named entities recognition, event detection). Particularly, the system targets NLP developers in a company, providing a stimulating environment where they can practice basic functionalities of ML applied to NLP.

The platform is delivered as a web application (see Figure 2) where multiple users are allowed to collaborate to the development of the same dataset. Through a graphical interface the user is guided to set up a project: this includes uploading unlabeled data, setting annotation categories, and setting re-training parameters (e.g. the frequency of the re-training). Then the user is presented with a document (e.g. a news story) selected from the unlabeled pool and is asked to revise the automatic annotation produced by the classifier with the model currently available. Once the user has confirmed the revisions, the system proposes a new document to be revised, on the basis of the AL selection procedures. At each moment the user can monitor the performance of the system on the task by consulting the learning curve and inspecting the content of the system memory (i.e. which errors are in the memory, how many times they have been considered, and whether the system considers them as solved or not).

We are not aware of any descriptions of similar platforms in the literature; while software packages for active learning do exist (for instance JCLAL) they are not integrated either with a graphical annotation tool or with an NLP pipeline. This is partly explained by the fact that in order to ensure replicability, research experiments on AL are typically performed on small annotated datasets and thus they do not need a real environment (with a real annotator). In concrete applications, on the other hand, more functionalities are needed as proposed in TEXTPRO-AL.

The TEXTPRO-AL platform is used in the context of four activities. The first is a collaboration with Euregio Srl[14] for developing a named entity recognition dataset for news in German from the South

[14] http://www.euregio.it

Tyrol area. In this case, the task and the categories used are standard, while the goal is to improve the performance on top of an existing dataset. The second experience is part of a research project on automatic analysis of live soccer commentaries in Italian (Minard et al., 2016b). In this case, the annotation categories (`player`, `goal`, etc.) were defined from scratch and the (non-expert) annotator was able to produce a dataset in seven working days. We also used the platform in a domain adaptation perspective for the annotation of named entities in tweets, where we annotated more than 2,000 tweets with the goal of adapting a system trained on news to social media texts (Minard et al., 2016a). Finally, we have been using TEXTPRO-AL for education purposes, to support an NLP introductory course.

6 Platform distribution

We currently offer the TEXTPRO-AL platform as an extension of the TextPro NLP pipeline (Pianta et al., 2008). TextPro is distributed under a dual licensing schema (i.e. free for research purposes, proprietary for commercial purposes). As we believe that domain adaptation is a major issue for extending the market of NLP applications, we are going to distribute the whole TEXTPRO-AL platform with the same dual schema adopted for TextPro.

Acknowledgments

This work has been partially supported by the EUCLIP (EUregio Cross LInguistic Project) project, under a collaboration between FBK and Euregio Srl.

References

David Cohn, Richard Ladner, and Alex Waibel. 1994. Improving generalization with active learning. In *Machine Learning*, pages 201–221.

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. Mt-equal: a toolkit for human assessment of machine translation output. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 120–123.

Anne-Lyse Minard, Mohammed R.H. Qwaider, and Bernardo Magnini. 2016a. FBK-NLP at NEEL-IT: Active Learning for Domain Adaptation. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*.

Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, and Mohammed R.H. Qwaider. 2016b. Semantic interpretation of events in live soccer commentaries. In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

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 *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 839–846, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Burr Settles. 2010. Active learning literature survey. Technical report.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Tomanek and Fredrik Olsson. 2009. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pages 45–48, Stroudsburg, PA, USA. Association for Computational Linguistics.