Visual Active Learning for Relation Extraction

Kairong Jiang
University of Arizona
jiangkairong@email.arizona.edu

Mihai Surdeanu
University of Arizona

1 Introduction

Relation extraction is an import part of information extraction (IE) where a classifier is trained to label the *relation* between a set of entity mentions in some text. It provides crucial information for later IE processes like disambiguation. For example, in the sentence "**Obama** *was born in* the **United States** just as he has always said.", the classifier should label relation *"BornIn"* with entity mention pair "Barack Obama" and "United States".

While supervised learning methods have been developed for relation extraction tasks, they typically require large amount of annotated training data to perform competitively. It is likely in real world problems that annotated data is limited or expensive to acquire. Therefore, it is beneficial to look for active learning methods that exploit a few informative annotated data and achieve reasonable performance while greatly reduce the work needed for human annotators.

On the other hand, existing active learning methods (Angeli et al., 2014; Fu and Grishman, 2013; Sun and Grishman, 2012) for relation extraction mainly focus on selecting the sampling strategies and improving the active learning model. While they have achieved notable improvements, the effectiveness and efficiency of human interactions are largely neglected in the aforementioned works. Human annotation is often simulated with fully-labeled data (Fu and Grishman, 2013; Sun and Grishman, 2012), or conducted using a listed multiple-choice view (Angeli et al., 2014).

In this paper, we present a relation extraction system implementing a distantly supervised model from Surdeanu et al. (2012) with a 2D scatter plot visual interface similar to Berger et al. (2018) for human annotators, and we conduct user studies to show that carefully designed and implemented visual interface can further improve the effectiveness and efficiency of active learning methods for relation extraction, primarily thanks to greater number of annotations that can be done with the same human effort. We also experiment with several sampling methods outlined in Angeli et al. (2014) and Berger et al. (2018) and explore the best sampling strategy suitable for the 2D scatter plot interface.

We are able to achieve following contributions through our studies:

- We present a 2D scatter plot visual interface for human annotations in relation extraction, which, despite lower accuracy, increases the number of annotations that can be made with the same human effort, hence improves the overall performance of the system.
- We experiment with different sampling methods and acquire understandings on the strong sides as well as draw backs of those methods, providing insights on sampling method selection with active learning using 2D scatter plot interface.

2 Related Work

Distant supervision has become a popular branch of approaches in relation extraction. It automatically generates labeled data by looking for the argument pair in the relation tables in a knowledge base (like Wikidata). Surdeanu et al. (2012) presents a multi-instance multi-label (MIML) model for distant supervision in relation extraction, addressing two major challenges in distant supervised models, namely incorrect labeling and multiple possible relations for one pair of entities. Still another problem exists where the knowledge base is often incomplete. Treating the missing knowledge as negative data will result in large

amount of false negative labels, which will hinder the model's performance. Min et al. (2013) addresses this problem by only generating positive and unlabeled data.

Active learning methods have also been developed during the years. Sun and Grishman (2012) presents a *co-testing* active learning model with local and global views. While local view is based on the context of a pair of entity, the global view classifies new examples based on the distributional similarity of the relation phrases. Fu and Grishman (2013) implements several improvements to the *co-testing* model including better initial setting and balancing imbalanced classifiers. Both of these works simulates user annotated data using fully-labeled data, thus neglecting the human interactions in the classification process. Also, these papers only test their models on a single dataset, providing limited experimental results.

Angeli et al. (2014) combines the `MIML` model presented by Surdeanu et al. (2012) with active learning. The paper present two criteria for selecting examples to annotate based on disagreement provided by QBC. Their experiments showed that combining the `MIML` model with annotated data yields improved results.

All of the previous active learning methods incorporate a list-based view presented to the human annotators. Bernard et al. (2018) conducts a thorough set of experiments comparing the effectiveness of traditional active learning strategies with visual-interface labeling methods, showing that visual-interface labeling can compete with traditional active learning methods in various tasks. Though in the paper uniformed sampling of example data is used as baseline while other sampling method may have better performance. Berger et al. (2018) presents a 2D-scatter-plot visual interface for human annotations of name-entity classification and conducts user-studies showing their visual interface out-performs traditional list-based active learning methods despite of noisy data. However, the paper doesn't compare different sampling methods and their visual interface approach can be extended to other information extraction tasks.

3 Baseline

We compare our model with the MIML_RE model presented by Surdeanu et al. (2012). The model is trained and tested on KBP 2010 and 2011 data (Ji et al., 2010, 2011), aligning the relations with both the knowledge based provided by the shared tasks and a snapshot of the English Wikipedia from June 2010. They use Stanford's CoreNLP package to identify entity mentions in text and they only consider entity mention candidates occurred in the same sentence. Table 3.1 shows the results of the MIML_RE model.

Table 1: Experimental results.

Model	Precision	Recall	F1
MIML_RE	19.5	30.7	23.8

The data and source code can be found at `https://github.com/Inlinebool/CSC585-Project`.

3.1 Error Analysis

As mentioned before, one major problem for distant-supervised models such as the MIML_RE model is that the knowledge base is often incomplete, resulting in a large number of false-negative examples. The model subsamples the negative examples at a rate of 5%, but the result can yet be improved by providing better negative examples, using human annotations. Another cause of error is that some relation pairs do not exist in the knowledge base, making it impossible for the model to learn those relations. This can also be mitigated by providing human annotations for the unknown relation pairs.

Other errors include system error inherited from CoreNLP and imperfect initialization.

References

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1556–1567.

Matt Berger, Ajay Nagesh, Joshua Levine, Mihai Surdeanu, and Hao Helen Zhang. 2018. Visual supervision in bootstrapped information extraction. In *To Be Published*.

Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2018. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*, 24(1):298–308.

Lisheng Fu and Ralph Grishman. 2013. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 692–698.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2010. Overview of the tac 2010 knowledge base population track. In *Proceedings of the Text Analytics Conference*.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the tac 2011 knowledge base population track. In *Proceedings of the Text Analytics Conference*.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

Ang Sun and Ralph Grishman. 2012. Active learning for relation type extension with local and global data views. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1105–1112. ACM.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465. Association for Computational Linguistics.