Progress Report

Hanchuan Li, Haichen Shen, Shengliang Xu and Congle Zhang
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{hanchuan,haichen,shengliang,clzhang}@cs.washington.edu

1 Literature Review

Manual annotation for NLP training data is well-known for its tedium and large amount of data. To generate a comprehensive annotated training set requires much human effort. Annotators are also prone to make mistakes during the long and tedious annotating process. Researchers are trying to address these problems by two means: 1) develop visualization tools to improve annotation efficiency as well as reduce the error rate in annotation; 2) adopt crowdsourcing to enable collaborative annotation that accelerates the process of annotation.

Most related in scope is (Yan and Webster, 2012) which provides a collaborative tool to assist annotators in tagging of complex Chinese and multilingual linguistic data. It visualizes a tree model that represents the complex relations across different linguistic elements to reduce the learning curve. Besides it proposes a web-based collaborative annotation approach to meet the large amount of data. Their tool focuses on a specific area —- complex multilingual linguistic data, whereas our work is trying to address how to generate a visualization model for general data sets.

Most related in scope is (Yan and Webster, 2012) which provides a collaborative tool to assist annotators in tagging of complex Chinese and multilingual linguistic data. It visualizes a tree model that represents the complex relations across different linguistic elements to reduce the learning curve. Besides it proposes a web-based collaborative annotation approach to meet the large amount of data. Their tool only focuses on a specific area that is complex multilingual linguistic data, whereas our work is trying to address how to generate a visualization model for general data sets.

Crowdsourcing now is recognized as a growing and promising approach in NLP. Many related works focus on conceptual study and formalization of crowdsourcing. For instance, (Quinn and Bederson, 2009) categorizes crowdsourcing into seven genres: Mechanized Labor, Game with a Purpose (GWAP), Widom of Crowds, Crowdsourcing, Dual-Purpose Work, Grand Serarch, Human-based Genetic Algorithms and Knowledge Collection from Volunteer Contributors. Other works, such as (Abekawa et al., 2010) and (Irvine and Klementiev, 2010), develops a specific tool and verifies the feasibility and benefit of crowdsourcing. Nevertheless, we seek to provide an intuitive visualization to lower the barrier to get started on crowdsourcing.

2 Project Plan

2.1 System overview

In this project, we aim to develop a visualized toolkit for crowd-sourcing NLP annotations. The target audience are normal people with little knowledge and patience. The toolkit would allow them to quickly label NLP datasets.

There are two key properties of our toolkit: firstly, annotators could interact with the data to understand them in a refresh way. Annotators label some examples and they expect immediate feedback from the toolkit. These feedbacks will help them understand the problem. Secondly, the toolkit should enable and encourage trial and errors. It would not assume any edits from the users as gold, but treat the edits as clues to better visualize the data to the annotator. When the annotator finishes a labeling task, he should be satisfied and confident with the overall outcome. For example, it is hard to distinguish whether “Jeff” is “Jeff Bilmes” or “Jeffery Heer”

when data points are seen individually. But if the toolkit could immediate show a big cluster {"Jeff, Jeff Bilmes, Jeffery Heer, Professor Heer"} after incorrectly merge two points, the annotators would have a good chance to fix it.

In this project, we would focus on two important kinds of NLP annotations: building trees (*e.g.* parsing) and clustering (*e.g.* coreference resolution). But we would keep in mind that the toolkit should be easily extensible to any NLP problems.

2.2 Milestones

System Brainstorming
All group members work on this together.

System Input Output Implementation:
Major Responsibility: Congle Zhang
Minor Responsibility: Shengliang Xu, Haichen Shen

System Graphic & Visualization Implementation:
Major Responsibility: Haichen Shen Shengliang Xu
Minor Responsibility: Hanchuan Li, Congle Zhang

System Layout Adjustment & User Evaluation Study:
Major Responsibility: Hanchuan Li
Minor Responsibility: Congle Zhang, Haichen Shen, Shehgliang Xu.

References

Takeshi Abekawa, Masao Utiyama, Eiichiro Sumita, and Kyo Kageura. 2010. Community-based construction of draft and final translation corpus through a translation hosting site minna no hon'yaku (mnh). In *LREC*. Citeseer.

Ann Irvine and Alexandre Klementiev. 2010. Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 108–113. Association for Computational Linguistics.

Alexander J Quinn and Benjamin B Bederson. 2009. A taxonomy of distributed human computation. *Human-Computer Interaction Lab Tech Report, University of Maryland*.

Hengbin Yan and Jonathan Webster. 2012. Collaborative annotation and visualization of functional and discourse structures. In *Proceedings of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing*, pages 366–374.