# 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

Many NLP tasks require large amount of high quality training data. Manual annotation for such training data is well-known for its tedium. 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) building specialized annotating tools to ease the annotating process in the hope of improving efficiency as well as reducing the error rates; 2) adopting crowdsourcing to scale up annotating.

**Specialized annotating tools**. Facing one of the biggest. common problems, many NLP researchers have developed a number of tools for annotating training corpora along the history of NLP research. At first, before the blossom of the web, tools are generally built as local programs such as the WordFreak linguistic annotation tool (Morton and LaCivita, 2003) and the UAM CorpusTool for text and image annotation (O'Donnell, 2008). These tools are very restricted because they cannot scale. Web-based annotation tools are developed later in order to scale up the annotating process, such as (Stührenberg et al., 2007). However these tools typically only use very basic HTML based techniques to provide very limitted visual aids for the annotating process. 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.

**Crowdsoursing in NLP**. Crowdsourcing (Howe, 2006) is a popular and fast growing research area. There have been a lot of studies on understanging what it is and what it can do. 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. It is generally convinced that crowdsourcing is of great beneficial if the tasks are easy to conduct by the workers and the tasks are independent.

Because of the high labor requirements in typical NLP training tasks, there also have been some work considing using crowdsoursing in many NLP tasks. For example, Grady *et al.*generated a data set on document relevance to search queries for information retrieval (Grady and Lease, 2010); Negri *et al.*built a cross-lingual textual corpora (Negri et al., 2011); Finin *et al.*collected simple named entity annotations using Amazon MTurk and CrowdFlower (Finin et al., 2010). Also there are some researchers observed the hardness of collecting high quality data and did some studies on improving that, such as (Hsueh et al., 2009)( how annotations should be selected to maximize quality), and (Lease, 2011) (quality control in crowdsoursing by machine learning).

Different from previous studies, we seek to improve crowdsoursing annotating quality by greatly lower the usability barrier through the proposed visualized toolkit rather than trying to cleaning up the data generated by the crowdsoursing process.

## 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.

### 2.2 System detailed design

We propose to build our system as a web application for collabrative annotation because we are targeting our toolkit as deployable by scalable crowsoursing systems. Based on this requirement, we plan to build the toolkit based on the D3 web visualization library.

#### 2.2.1 Input Output design

As a collaborative web application, the input/output data must be sharable by different annotators. We plan to use the data storage service provided by the Google app engine.

#### 2.2.2 Data representation design

By survying a lot of existing NLP tasks, we decide to focus on two types of annotating data. 1) Given an article or a webpage and a list of entities represented by words or phrases, where the entities appear in the article, annotate the entities; 2) Given a list of sentences, paragraphs or articles, directly annotate them.

#### 2.2.3 Task visualization design

In this project, we would focus on two important kinds of NLP annotations: building trees (*e.g.* parsing) and clustering (*e.g.* coreference resolution).

We propose to build a new D3 tree plugin for conducting the tree building task by visualizing the in-building trees. The users can directly operate on the visualized tree to complete the whole annoting process. In addition to the tree builidng from a set of unstructured data points, we also plan to support tree evolving, i.e. building other trees from an existing tree. This feature is applicable to many cross-lingual tasks such as mapping a semantic tree of an English sentence to the tree of the translated Chinese sentence.

For the clustering task, we propse to do it by building a graph based clustering plugin on D3. The users can directly operate on the clustering graph to finish the clustering annotating processing.

### 2.3 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.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pages 27–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Human Computation*.

Thomas Morton and Jeremy LaCivita. 2003. Wordfreak: An open tool for linguistic annotation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4*, NAACL-Demonstrations '03, pages 17–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mick O'Donnell. 2008. Demonstration of the uam corpustool for text and image annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 13–16, Stroudsburg, PA, USA. 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*.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Alexander Mehler, and Irene Cramer. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 140–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.