

A Visualized Toolkit for Crowdsourcing NLP Annotations

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 (?) 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 (?) 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, (?) categorizes crowdsourcing into seven genres: Mechanized Labor, Game with a Purpose (GWAP), Wisdom of Crowds, Crowdsourcing, Dual-Purpose Work, Grand Search, Human-based Genetic Algorithms and Knowledge Collection from Volunteer Contributors. Other works, such as (?) and (?), develop a specific tool and verify 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

References