Distributed Research Experience for Undergraduates 2009

Final Report

Kriti Godey

I participated in the the Distributed Research Experience for Undergraduates (DREU) program (origanised by the Computing Research Association) for ten weeks in Summer 2009. This is the final report of my research.

My research mentor was Dr. Yi Zhang, Assistant Professor of the University of Santa Cruz College of Engineering. I was working in her tecWAVE development lab under the guidance of Jiazhong Nie, one of Prof. Zhang's first-year doctoral students. I also attended weekly meetings of the IRKM (Information Retrieval and Knowledge Management) seminar, led by Prof. Zhang.

The tecWAVE Project

tecWAVE stands for **te**aching with **c**omputers: **W**ord **A**nnotations for **V**ocabulary **E**ducation. It is designed to assist students with their vocabulary by annotating words that might pose difficulty. It is especially geared towards English language learners. In its current stage, the WaveMachine focuses on middle school science texts. These texts are input into the system and certain words are highlighted, allowing students to click on them. These words are automatically highlighted by the application, based on algorithms depending on many factors including age of the user, grade level, reading level and native language. When the user clicks on a word, a window pops up with definitions of the word and pictures associated with the word. For effective comprehension, these definitions are present in both the student's native language and English.

Roles

There are several ways the tecWAVE application can be used. These are represented via “roles” such as *Researcher* and *Teacher.* Researchers are only allowed to extract generic demographic data regarding how the application is utilised. Researcher privileges are granted by the tecWAVE team by application.

The *Teacher* role provides tools for a more direct application of tecWAVE. Teachers have the many capabilities, including tools to:

1. Register a group of students in bulk
2. Register individual students
3. Register their colleagues and peers.
4. Generate reports about their students' progress.
5. Provide feedback for definitions or images.

The *Student* role is an actual user of the WaveMachine for vocabulary education.

Features

The following is a list of features that the WaveMachine possesses:

1. It will automatically annotate words that a student is unfamiliar with, based on grade/reading level using artificial intelligence, especially natural language processing techniques such as:
    - automatically disambiguating the annotations based on context (for example, if a word has multiple definitions, selecting the correct one based on its place in the text.)
    - automatically generating pictures, pronunciation keys and dictionary entries for the selected word and context.
    - Automatically providing an annotation in the user's native language, when appropriate.

2. Based on explicit and implicit feedback from the user's previous usage of the WaveMachine, it will personalise the annotations individually, figuring out which words should be highlighted and annotated.

3. The application tracks the history of students' learning and progress by their use of clicks through activities. This assists in refining the system so that it works more effectively in the classroom.

4. The teacher is integrally involved in the instruction and use of the WaveMachine among his or her students. Teacher input allows the application to accept social annotations through Web 2.0 techniques to further improve the accuracy of the algorithm's word-picking system. In addition, the teacher can use the application to generate a variety of reports – summarising the progress of an entire class, as well as individual student reports. This information provides valuble instructional feedback for all aspects of the curriculum.

5. Using Web 2.0 techniques, the input of the students' parents can also be solicited to allow the system to accept social annotations. This will also help improve the quality of the application's further annotations.

My Research

The tecWAVE team consisted of around twelve people, each focussing on improving a different aspect of the program, both in engineering and educationally. I was working with Jiazhong Nie, who was responsible for the text definition module of the project. This module provides definitions for English words in three languages - English, Chinese and Spanish. For words with multiple meanings, definitions are chosen according to the context in which they occur. This disambiguation process is performed according to a supervised machine learning approach.

Jiazhong had two major responsibilities – translation and word sense disambiguation. My focus was almost entirely on the word sense disambiguation algorithms. For each ambiguous word, a log regression classifier is trained based on a corpus of human tagged text. This trained algorithm is then applied to the text needing annotation. We were trying to enrich the algorithm by the technique of multi-task learning, in which various kinds of information are captured and utilised in the disambiguation process.

My first task was to optimise large data files of matrices into formats that Jiazhong could use in his experiments. The matrices represented data retrieved from various corpuses of text (computed

before my research began.) I wrote Python programs that modified these matrices into a number of different formats, including randomly generating certain rows and changing specific parameters of a single column. I had never used Python before, so I was learning a new language as well.

My next task was to extract title and article information from the gigantic Reuters News Corpus. This file was formatted in XML, and I had to extract data in a readable format, with no special characters. I also had to format the resulting files in a easily indexible manner. Once these files were ready, I had to modify a pre-written script to read these files and run on multiple simultaneous processes, and run upwards of fifty experiments using different parameters (specified in the script.) Jiazhong looked through these results to choose how to modify his algorithm, and concluded that this form of multi-task learning did not produce a worthwhile result.

Most of my research consisted of conducting these experiments and writing scripts and programs. I also used data from SemCorp and SemEval and tried to run it through the Berkeley Parser and make a semantic tree from the data. I had issues with the parser, though, and that would not run properly. I managed to fix it eventually with Jiazhong's help and with an e-mail to the developer of the parser. That was an interesting experience – using an extremely specialised tool and attempting to troubleshoot it.

Additional Work

Apart from my research at the tecWAVE lab, I did some additional research work and used my free time to learn some new things. I have listed some of them below:

1. I attended weekly meetings of the Information Retrieval and Knowledge Management seminar led by Prof. Zhang. This seminar consisted of many of her graduate students working on a variety of different projects. Every week, they discussed their own research and the progress made, and/or presented summaries of papers from conferences they had been to, or papers relevant to their own research experience. This was extremely informational and I learned many new techniques and mathematical tools. It was also refreshing to get perspectives on research beyond my own project.

2. At one of these IRKM meetings, I was asked to present the paper “Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition” by Arnold, Nallipatti and Cohen from Carnegie Mellon University. This paper was about a technique for supervised transfer learning in the recognition of personal name mentions. This technique was very similar to what Jiazhong and I were using in our algorithm development. I gave a fifteen minute presentation to the seminar summarising this paper, including the mathematics involved and future work to be done.

3. I attended a day long seminar at the NASA Ames Research Centre, where two separate research groups that Prof. Zhang was affiliated with met up to discuss papers presented at the SIGIR 2009 conference. That was extremely fascinating, especially because the setting was at a NASA research centre.

4. In my spare time, I wrote Python programs in an effort to learn more of the language. I worked my way through a set of mathematical problems from the website “Project Euler” I also

installed and experimented with Django, a content management framework for Python. In addition, I constructed a research and personal website for Prof. Shastry, a professor in UCSC's Physics department.

References

1. *tecWAVE Overview;* http://tecwave.soe.ucsc.edu/over.php – retrieved on January 7, 2010.
2. Arnold, Cohen, Nallipati; *Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition*, Carnegie Mellon University.