Project Proposal

Yucen Li (Lily)

Project Title

Selection of Languages For Cross-lingual Parsing

Project Web Page

The updates to my project will be posted to https://www.andrew.cmu.edu/user/yucenl

Project Description

For my project, I plan to work with Professors Lori Levin and David Mortensen from the LTI Department. The project is based on Universal Dependencies, which is an effort to represent the dependency grammars of different languages throughout the world as consistently as possible, with similar tags and relationships between words. These treebanks can be used for cross-lingual training, which has been shown to increase the accuracy of parsers depending on the relationship between the languages selected. In my research, I will analyze different linguistic properties of languages, such as common typological features such as subject object order etc, to form hypothesis regarding which features are most important in predicting the success of cross-lingual training.

After the predicted hypothesis have been formed, it is important to select the relevant cross-lingual embedding models to ensure that the different in results are mostly due to linguistic differences between the languages rather than a mistranslation.

Another confounding variable in the accuracy of the parser is the quality of the Universal Dependencies. Many of the treebanks in the Universal Dependencies have been automatically converted from existing treebanks with different standards. Therefore, it is possible that the treebanks are inconsistent, and the UD treebanks may not be representative of the language. In order to reduce this variable, I will analyze the quality of the treebanks to see how well they conform to the UD standards.

Because dependency parsing and grammar is heavily used in many tasks in natural language processing such as coreference resolution, question answering, information extraction, etc, develop-
ing better parsers is extremely beneficial to many tasks. Crosslingual is particularly helpful for low resource languages, as they may have very limited amounts of training data. Therefore, finding the best languages to optimize crosslingual parsing is beneficial for many language-related tasks.

**Project Goals**

For my project, I will evaluate the success of my project based on the accuracy of the parsers compared to existing crosslingual or monolingual parsers.

A 75% goal would be to develop more accurate parsers for a specific set of languages. For example, we can look inside the subset of Romance languages to see which features are most indicative of cross-lingual success.

The 100% goal is to find generalized rules for all languages which determine which other languages should be selected for cross-training. For each language, I would like to find a systematic way to select the languages for cross-lingual parsing.

The 125% goal for the project would be to find good crosslingual results for low-resource languages, or possibly obtain good parsers which were trained only on treebanks from other languages.

**Project Milestones**

**End of Semester**

I will continue to read research papers to see what current work has already been done in cross lingual parsing. I also want to become familiar working with Dynet, which is a dynamic neural network which is often used to build syntax parsing systems..

**January 31**

Be completely set up with the tools and cross-lingual training models. I will then try to reproduce the results from previous crosslingual parsing models.

**February 14**

Spend time identifying linguistic similarities between languages which may be predictive of cross-lingual performance. Formal hypothesis should be complete.
February 28

Experiment with different architectures for language embedding depending on the hypothesis formed. I will also develop the basic models to train and test on.

March 21

Analyze results of experiments and modify hypothesis. Also factor in quality of treebanks to see if inconsistency within treebanks affects cross lingual parsing ability.

April 4

Perform more experiments based on modified hypothesis depending on previous experimental data. Analyze for linguistic reasoning behind results.

April 18

Compare results of the parser with existing cross lingual or monolingual parser approaches.

May 2

Complete analysis of the work for the semester and present.

Literature Search

I am currently reading Many Languages One Parser [1] which introduces a crosslingual parser. I plan on reading more papers to have a better idea of what research has already been done in this area. I also want to learn more about crosslingual embedding. I have also been reading the documentation for the Universal Dependencies so I know how the treebanks are structured.

Resources Needed

I will need to setup DyNet, which is available on GitHub. For research next semester, I will also have access to a server within LTI so I will be able to train and test my parser.
References