Natural Language Processing Term Project Proposal

Group: Chen Chen, Weihang Song
Wenzhao Zang, Min Guo

Option 1. Build a system that takes a news article, extracts its main idea and key information, and presents them to the user in the format of a summary paragraph or bullet points.

(1) We'll use labeled news articles as the training set (including but not limited to WSJ), and follow the workflow consisting of POS tagger, chuck tagger, name tagger etc..

(2) Given that training sets on summary extraction might not be available, we need to produce some rules on summary generation, such as where the key information is likely to occur inside an article, or how to glue different information pieces together to create a summarizing sentence. We might start with a small sample and program some simple rules to see the quality of the summary paragraph, and then refine the rules, expand the samples or improve the algorithm accordingly. It'll be a long trial-and-error process.

(3) The end results might be organized following a template. For example, the user will be able to see something like: "This article is on the topic of xxx, first it says xxx is xxx… second it says…". The user can then decide whether to read the article in depth for more details, or be contented with such high-level information and move on.

Challenges:

(1) Training sets on news articles other than WSJ might not be available, but are very important, and preferably there should be complete information on POS, chuck, and name tagging.

(2) Given the lack of training sets for the second phase, the alternative rule-based design might not work or yield low accuracy.

Option 2. Chinese Natural Language Processing and Speech Processing

Build a Chinese Natural Language Processing system including word segmentation, Part-of-Speech tagging and Named Entity Recognition.

Chinese is standardly written without spaces between words. It means segmentation process for Chinese NLP is important and indispensable. Some papers already talked about how to do it and we might follow those advice to use linear-chain conditional random filed (CRF) model, which treats word segmentation as a binary decision task to do words segmentation.

In POS step, our system takes word-segmented Chinese text as input and assigns a part of speech to each word (and other tokens), such as a noun or a verb. Since we already implement POS process for English, we might use the same method first and then optimize our process.

For Name Entity Recognition, we will do the similar thing in POS step.

Challenges:

(1) Lack of training sets for Chinese words.

(2) Chinese words segmentation.

Option3. Extracting key information of Chinese article

If possible, build an extracting key information system for Chinese articles. Combining part1 and part 2, abstract the main idea of Chinese articles given by users.