Homework 4: Project-Related Paper Report

Guangyu Lin, EID: gl8429

April 21, 2016

1 Summary

I will write a paper review for Xin Li and Dan Roth's paper, which is Learning Question Classifiers: The Role of Semantic Information. (1)

Open-domain question answering and story comprehension have become important directions.(Question answering task is to seek an accurate and concise answer to a free-form factual question from a large collection of text data) A question answering task is different from the common information retrieval task done by search engines. It is more difficult because the textual resources are more confined and the target text is less likely to exactly match text in the questions. Therefore we need some advanced natural language techniques rather than key term extraction and expansion.

In order to study several possible semantic information sources and their contribution to classification, they also compare four types of semantic information sources that differ in their granularity, the way they acquired and their size, which are,

1. Automatically acquired named entity categories
2. Word senses in WordNet 1.7 (3)
3. Manually constructed word lists related to specific categories of interest (4)
4. Automatically generated semantically similar word lists

Li and Dan's strategy is to augment the questions with syntactic and semantic analysis, as well as external semantic knowledge, as input to the text classifier. It can also be viewed as a case study in applying semantic information in text classification. Their experimental study focuses on,

1. Testing the performance of the classifier in classifying questions into coarse and fine classes.
2. Comparing the contribution of different syntactic and semantic features to the classification quality.

In the experiment, Li and Dan observe that classification accuracies over 1,000 TREC (5) question reach 92.5% for 6 coarse(ABBEVIATION, DESCRIPTION, ENTITY, HUMAN, LOCATION, NUMERIC, VALUE) classes and 89.3% for 50 fine-grained classes. An error reduction of 28.7% can be achieved when semantic features are incorporated into fine-grained classification.

Generally, in this paper, Li and Roth shows that we need to understand the question to a level that allows determining some of the constraints the question imposes on a possible answer, which helps us respond correctly to a free form factual question. These constraints may include a semantic classification of the sought after answer and may even suggest using different strategies when looking for and verifying a candidate answer. A semantic classification is obviously a machine learning question. Li and Roth developed a hierarchical classifier that classifiers questions into fine-grained classes. They also perform a systematic study of the use of semantic information sources in natural language classification tasks. And their experiment results show that augmenting the input of the classifier with appropriate semantic category information results in significant improvements to classification accuracy.

2 Improvements

Though Li and Don compare the experiment results among Name Entities, WordNet Senses, Class-Specific Related Words and Distributional Similarity Based Categories, we still have much work to do. As State-of-the-art Named Entity Recognition(NER) systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95% (6). Additionally, NER is also a context sensitive semantic analysis of words. Therefore, we would like to compare NER with Li and Don's current features. The example below will give us an idea of the useful of combining NER.

"How did serfdom develop in and then leave Russia ?"

Li and Don labeled this sentence as **DESC:manner** , however, we also know **Russia** is an important information in this sentence when we answer this question. NER can help us retrieve this kind of information. Therefore, we will deep dive Li and Don's Question Hierarchy model by combining words label, NER. However, our NER label is not a traditional label as previous label. We would like to give each word a label. Word2Vector is another approach we will apply. We would like to use Google's Word2Vector to train our corpus, then concatenate the 1 bit NER label with each word's vector. For example, if a word is transferred to a 300 length vector, we will add it to 301 length vector by adding NER label. Then we will train convolutional neural networks for training the classification. (7), which will be discussed in my partner's report in detail.

Evaluation

We will use TREC (1) dataset to evaluate our natural language classifier. This dataset includes several corpus from 1,000 questions to 5,500 questions and 500 test questions. After labelling these data by combining NER's label, we will train another classifier by convolutional neural networks. (7) In the end, we will compare our result will Li and Dan' s state of art accuracy. Hopefully, we will get a better performance that Li and Dan.

References

[1] Li, Xin, and Dan Roth. "Learning question classifiers: the role of semantic information." Natural Language Engineering 12.3 (2006): 229-249.

[2] https://www.cs.utexas.edu/ mooney/cs388/hw3.html

[3] Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.

[4] Pantel, Patrick, and Dekang Lin. "Discovering word senses from text." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[5] Voorhees, Ellen M., and Chris Buckley. "The effect of topic set size on retrieval experiment error." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.

[6] Marsh, E., and D. Perzanowski. "Muc-7 evaluation of information extraction technology: Overview of results." Seventh Message Understanding Conference (MUC-7). 1998.

[7] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).