NLP report

Ge Gao gg24984

gegao1118@utexas.edu

1. Introduction

Automatic question answering is an important information retrieval task in natural language processing and it is more challenging than purely searching task because it needs to "understand" what is asked before searching for answers. And to make machines "understand" a question is as challenging as it sounds like. One method that could help solve this problem is called Question Classification(QC). Question Classification(QC) is a task that given a question, some classifier will map this question to one of k limited classes, so that some semantic constraint will be put on this question, which provides machine with some basic information about the question which could hopefully help machine "understand" questions.

In 2000's TREC competition, participants were requested to develop a system to classify English questions based on some question categories. And after that, many remarkable results have been achieved. One of them is Li and Roth's work. Li and Roth develops a hierarchical classifier based on SNoW learning architecture(Carlson et al, 1999; Roth, 1998) to solve the question classification task.

With the development of deep learning, machine learning models have been applied to NLP tasks and it turns out that many of these models significantly improved performance of original models without deep learning methods. One example is Yoon Kim's work that applies convolutional neural networks(CNN) to sentence classification.

Since named entity is an important part of questions, especially when number of question types is large, but previous work did not make much use of the semantic information of a named entity. For example the named entity might be either a location or a person. This work makes use of Stanford's Named Entity Recognizer(NER) and adds extra labels to vectors generated by word2vec. Then we train convolutional neural network model with these new vectors. Our goal is to improve the performance of the CNN model for question classification, especially for questions related to named entity semantic meanings.

TODO – introduction conclusion

2. Related Work

2.1. Question classifiers

In the "Learning Question Classifier" paper(Li and Roth, 2002), in terms of the question classification task, the number of question types(k) could be either six or fifty depending on different classification criteria. Li and Roth developed a machine learning method to classify questionsïijŇ which is guided by a layered semantic hierarchy of answer types. They made use of a sequence of two simple classifiers to do the classification. The first classifies questions into coarse classes (six in total) and the second classifies questions into fine classes(fifty in total), which is dependent on the first one. Here is the structure of the hierarchical classifier by Li and Roth.

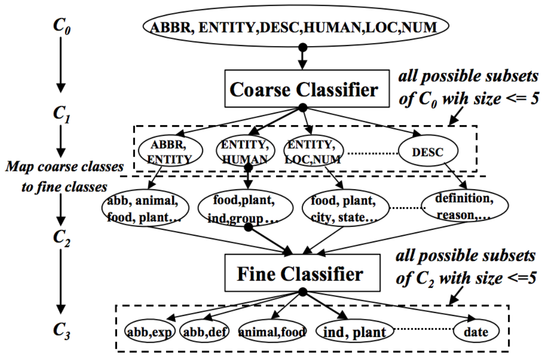
C_0
ABBR, ENTITY,DESC,HUMAN,LOC,NUM
Coarse Classifier
all possible subsets of C_0 wih size <= 5
C_1
ABBR, ENTITY
ENTITY, HUMAN
ENTITY, LOC,NUM
DESC
Map coarse classes to fine classes
abb, animal, food, plant…
food,plant, ind,group…
food, plant, city, state…
definition, reason,…
C_2
Fine Classifier
all possible subsets of C_2 with size <=5
C_3
abb,exp
abb,def
animal,food
ind, plant
date

Figure1: The hierarchical classifier

Figure 1 shows the basic classification process by classifier developed by Li and Roth. A question will always be processed along a top-down path to be classified. And after the classification process, question type label(s) will be attached to the question.

In Li and Roth's classifier, each question is analyzed as a list of features so that they could be trained and tested for learning.They extracted some primitive features like words, pos tags and chunks (Abney, 1991), named entity as well as some semantically related words. Based on these features, they compose and make some more complex features. Also, they make a semantically related word list for each of most question types. For example, "far" is in the semantically related words of "distance" so that if there is an occurrence of "far", and then the sensor for "distance" will be activated and the feature will be extracted.

A point that is worth pointing out is that there might be ambiguity for some specific questions. For example, a question like "What do bats eat" could either be classified to belong to food type or animal type, and both them make much sense. To solve this, Li and Roth allow multiple type labels to be attached to a single question.

2.2. CNN on sentences classification

Convolutional neural networks(CNN) model is a deep learning method which has achieved remarkable results in many fields. A CNN model was developed by Yoon Kim for sentence classification which accepts word vectors as input and generates type labels as output.

Convolutional neural networks was originally invented for tasks in computer vision fields and was proved to be also effective in many NLP tasks. In Kim' model, they train one layer of filter on top of word vectors generated by word2vec developed by Google(Mikolov et al 2013). In Kim's model, there are "static" and "nonstatic" models for model variation, where "static" means that the vectors are directly from word2vec and for "nonstatic", those vectors will also be tuned for each data set.

3. baseline of Kim's model

Because Kim's model that is available from his GitHub is for Google news, so we made some modifications based on his code and did the replicating experiment for TREC 10.