The University of Melbourne

Department of Computing and Information Systems

Student Number and Name: Jehle Yu

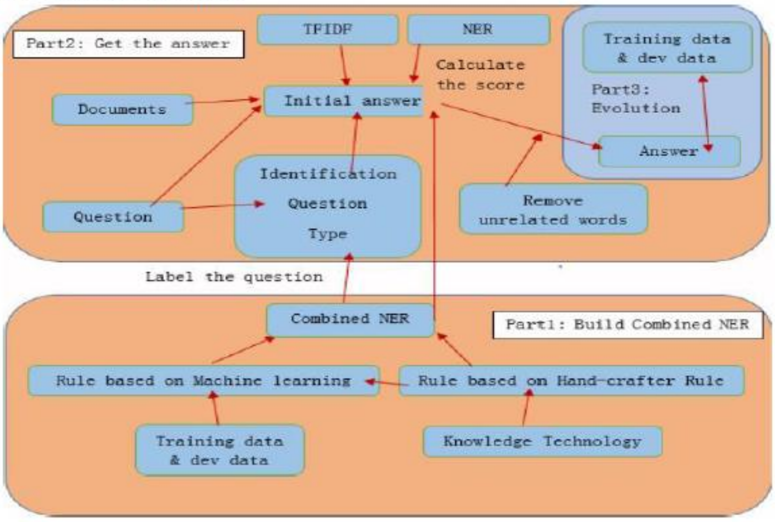
1. Background

Question-Answering (QA) is an example of such since answers are frequently named entities in agreement with the semantic category expected by a given question (Mendes, A.C., Coheur, L., and Lobo, P.V., 2010). So, in this project, I want to use the named entity recognition (NER) to build the QA model. In here, I combine the hand-crafted rule, machine learning, the Stanford NER version 3.9.1 and spacy NER as my named entity recognizer. The TFIDF is used to find the sentence where the answer is.

2. Introduction

This report will introduce the data preprocessing and how to build a basic model in section 3. And then, it introduction how to optimize the model in section 4. Finally, it will have the conclusion in section 5 and show the future work in section 6.

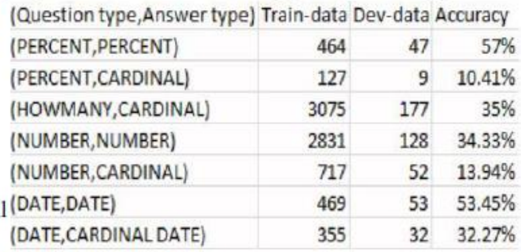
3. Method Description



Graph1: method flow chart

3.1 Build Combined NER (Part1)

In here, I use the Rule based on the Hand-crafted rule and spacy NER to find all the entity pair (question type, answer type) between question and answer in training data and then use the model to label the question in dev data. The rule based on the hand-crafter rule is based on our life experience, for example, if the question has the word 'number', 'frequency', 'how many', the answer almost NUMBER type, so I set the question type to NUMBER type. If a question can't be judged by our knowledge, it will be labeled 'OTHER'. The answer set will be judged by the spacy NER. If the answer can't be labeled, it will be labeled 'OTHER' too. This table is the accuracy of prediction

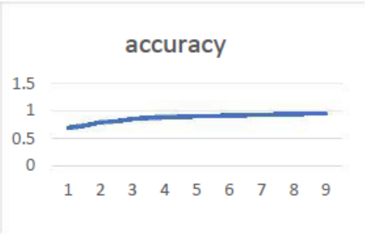


(Question type,Answer type)	Train-data	Dev-data	Accuracy
(PERCENT,PERCENT)	464	47	57%
(PERCENT,CARDINAL)	127	9	10.41%
(HOWMANY,CARDINAL)	3075	177	35%
(NUMBER,NUMBER)	2831	128	34.33%
(NUMBER,CARDINAL)	717	52	13.94%
(DATE,DATE)	469	53	53.45%
(DATE,CARDINAL DATE)	355	32	32.27%

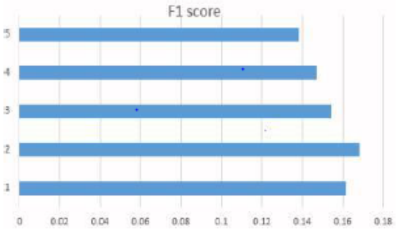
question type in dev data. This shows that there is a high probability of finding the answer type through this method. And we can find that still some accuracy under 30%, it is because of spacy sometimes can't define the answer type. The result of rule based on Machine learning will store as a dictionary, thus the combined NER will first use the hand-crafter rule to judge label and then use the NER because of the question in English is generally a fixed vocabulary, which can be summed up from our experience.

3.2 Get the answer (Part2)

I use the TF-IDF to calculate and get the paragraph where the answer is and then use the combined NER and Question type (method in 3.1) to find the answer. In here, I use the dev data to test the accuracy of finding the answer segment from top N paragraph with the highest score. From graph2, we can know that the use top 3 paragraphs already can get good results. Considering the running time, I choose to return top 3 paragraphs. And then I calculate the cosine similarity between question and each sentence in the top 3 paragraphs, hoping to find the sentence where the answer is and then find the answer. The answer will be selected from the sentence with the highest value. First choose the word with the same type of question prediction, if not, return this sentence. In here, I want to sacrifice recall to improve accuracy through use N possible answers as one answer. Graph3 shows the result that use the dev data to compare F1 score for use 1-5 possible answers as one answer, so I choose to use two answers as one answer. And now, I have an initial QA model, it will return short words or one sentence.



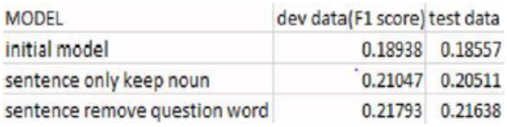
Gragh2: Choose Top N paragraphs



Gragh3: Choose Top N possible answers

4. Comparative assessment

In order to improve the recall, I want to reduce the number of answers and remove unrelated words. The F1 score of the initial QA model to return answer according to the dev data's question is about 19% and the F1 score of kaggle evolution the answer of test data is about 18.5%. By analyzing the question and answer of train data, I found that most of the questions are quest a nouns form answer. So I optimized the model with only keep the nouns form words in the sentence as part of the answer when the model returns a sentence because of it can't find a word has the same type as question expected. And there is about 2% increase in optimized results of test data. Moreover, I find the answer in train data almost don't have the words that have appeared in the question, so I remove the question word from the answer and the new result of test data has 1% increase. But when I check the answer set, I found some answer is not fluent because of I only keep the noun before, so I add the word near the noun word in the answer and make it more like a human answer. And the F1 score of test data become 0.22818.



MODEL	dev data(F1 score)	test data
initial model	0.18938	0.18557
sentence only keep noun	0.21047	0.20511
sentence remove question word	0.21793	0.21638

5. Conclusion

According to this project, I found that the application of NER on the QA system is not very good. It performs well on the question concerning number, person, and place while performs a little poorly on other types of issues. That's because of it can't predict the exact position of the answer in the sentence and can't handle unknown types of questions and answers.

6. Further work

In the further work, I will use the RNN and LSTM to training model to predict the start and end position of the answer. And build a question classifier to reduce the number of unknown type question and strive to divide all question into the different category.

Reference

Mendes, A.C., Coheur, L. and Lobo, P.V., 2010, May. Named Entity Recognition in Questions: Towards a Golden Collection. In *LREC*.

Narayanan, S. and Harabagiu, S., 2004, August. Question answering based on semantic structures. In Proceedings of the 20th international conference on Computational Linguistics (p. 693). Association for Computational Linguistics.

Pasca, M.A. and Harabagiu, S.M., 2001, September. High performance question/answering. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 366-374). ACM.