COMP90042 Project 2018: Question Answering

Username: shuaiw6

Kaggle team name: Shuai Wang

1 Introduction

The challenge of the project is to build a Question Answering (QA) system. More specifically, given a question and a document, the goal is to identify the answer to that question in the corresponding document.

2 Techniques

My QA system uses a typical architecture consisting of three components linked sequentially:

- Question Processing (QP), which identifies the type of the input question.
- Passage Retrieval (PR), which retrieves a small number of relevant passages from a document.
- Answer Extraction (AE), which extracts best answer from the previously retrieved passages.

This section describes all these components.

2.1 Question Processing

The QP component detects the type of the input question by mapping them into a two-level taxonomy consisting of 6 question types and 53 subtypes:

Type	Subtype
ABBREVIATION	abbreviation, expression
DESCRIPTION	definition, description, manner, reason
ENTITY	animal, body, colour, creative, currency, disease/medicine, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUMAN	description, group, individual, title
LOCATION	city, country, mountain, other, state
NUMERIC	code, count, date, distance, money, order, other, period, percent, temp, speed, size, weight

Figure 1 Question types from Li and Roth (2002) [1]

For each question, we assign such a two-level question type. For example:

- HUM: ind Who developed the Macintosh computer?
- DESC: def What are spider veins?
- NUM: date What year did Canada join the United Nations?

I use SVM based linear classifier to build a model on TREC corpus [2] to classify a given question to a correct question type. Following features are extracted to train our model:

- Token words (All tokens)
- Token lemma
- POS tags (Part of speech tags)
- WH word type (WH words represent what, which, when, where, who, how, why)

- Key words (e.g. Which team do you like? The key word is “team”)

There are 5,953 questions assigned with a two-level question types in TREC corpus. I split the question set into training set with 5453 questions and testing set with 500 questions. Train our model on the training set and evaluate it on testing set. The following figure shows the results:

Type	Gold	Model	Correct	Precision	Recall	F1-Score
All	500	500	363	0.726	0.726	0.726
ABBR	9	7	7	1.000	0.778	0.875
DESC	138	243	135	0.556	0.978	0.709
ENTY	94	60	40	0.667	0.426	0.519
HUM	65	62	55	0.887	0.846	0.866
LOC	81	52	51	0.981	0.630	0.767
NUM	113	76	75	0.987	0.664	0.794

Figure 2 Question classifier evaluation matric (for simplicity, I just showed the first-level question type)

The goal of question classification is to help answer extraction (We talk later how it can do that).

2.2 Passage Retrieval

The PR component is to retrieve a specific sentence in document that might contain the answer to the given question. The idea of PR is based on information retrieval. The question can be regarded as a query. The document that contains the answer to that question consists of several paragraphs. Each of them can be segmented into sentences. Thus, we can view a sentence as a document in IR. Now we just need to retrieve the sentence that is most relevant to the given question, which might contain our answer. Here, passage means sentence.

I use TF-IDF vector space model to implement passage retrieval. Train our model on training set and evaluate it on development dataset. One important thing is that the paragraph that contains retrieved sentence by our model can be viewed our prediction. If the paragraph is the right paragraph that contains right answer, then this prediction is right. In this way, our model is evaluated. The accuracy follows:

Gold	Model	Accuracy
3097	1997	0.645

Figure 3 Passage retrieval model accuracy

In my implementation, PR model retrieves three sentences to a question. This is because the most relevant sentence does not represent it contains the correct answer. So once a sentence cannot provide answer, check next two sentences. This will improve the performance of PR.

2.3 Answer Extraction

The final stage of AE is to extract a specific answer from the retrieved sentences. How to do that? Basically, I use question types and named entities:

- Question type mapping to named entity.

First, we use our question classifier to assign a question type to a question, and then use Spacy NER [3] to find all entities in retrieved sentences. There are some mappings from question types to named entities (See mapping tables in appendix). We can use those mappings to extract our answers. For example:

Question	Who was the UN under-secretary in 2011?
Question Type	HUM_ind
Sentences	[u'In 2011, UN under-secretary general Achim Steiner said: "The continuing growth in this core segment of the green economy is not happening by chance.', u'The group is to be co-chaired by Kandeh Yumkella, the chair of UN Energy and director general of the UN Industrial Development Organisation, and Charles Holliday, chairman of Bank of America".', u'Renewables producing electricity accounted for almost half of the 208 GW of capacity added globally during 2011.']
Named Entity	[(u'2011', u'DATE'), (u'UN', u'ORG'), (u'Achim Steiner', u'PERSON')]
Answer	achim steiner

Figure 4 Example of question type mapping to named entity

For the question, our question classifier classifies it into HUM_ind type, which means we need to find an individual human in retrieved sentences. Spacy NER recognizes all entities in those sentences. In the first sentence, we can find a "PERSON" entity which is "Achim Steiner". Thus, it can be our answer to the question. This is how we use question type and NER to extract answer.

3 Analysis

In this section, let's discuss some challenge I've encountered and explore the reason.

Challenge 1: Question classifier performs bad on "What" questions

In training set, there are 16911 "What" questions, 4384 "Who" questions, 4303 "How" questions, 3217 "When" questions, 2219 "which" questions, 1614 "where" questions, and other type of questions. Our question classifier can classify "Who", "How much", "How many", "When", "Where" questions into relatively accurate question types. But, it's hard to identify the types of some "How", "Which" and other questions. In figure 2, we can see that even through the first level type can be identified with higher precision and recall, but for second level types (subtypes) it performs bad. The reason is our TREC dataset only have 5953 labelled questions, it's not large enough.

Challenge 2: Passage retrieval model got low-accuracy performance

Our passage retrieval model got the accuracy of 0.645, which is not good enough. If the model cannot return right paragraph(sentences), it goes into wrong paragraph and try to get trivial answer. The performance of our question answering system directly depends on the choice of right paragraph. The main reason is that a relevant paragraph or sentence does not mean it contains the answer.

Challenge 3: There is no one-to-one mapping from question type to named entity

As you see from mapping table (see in appendix), we have 53 question types, but only 18 types of named entity. It means that for many of question types, there is no mapped named entity. So, we cannot do answer extraction. By the way, in my implementation, if there is such mapping, just return the whole most relevant sentence as answer. This is the main reason that I got a lower Kaggle score.

4 Kaggle Score

My Kaggle score is 0.166

5 Next

For next steps to improve the performance of my Question Answering system. I can do the following things:

- Combine question types with answer-extraction patterns to extract answer.
- Build a classifier to ranked candidate answers

- Extend the number of named entities that Spacy NER can recognize
- Use larger dataset to train question classifier

6 References

[1] Li, X. and Roth, D. (2002). Learning question classifiers. In *COLING-02*, pp. 556–562.

[2] http://cogcomp.org/Data/QA/QC/train_1000.label

[3] https://spacy.io/usage/linguistic-features#section-named-entities

7 Appendix

Question Type – Named Entity mapping table.

Question Type	Named Entity	Question Type	Named Entity	Question Type	Named Entity
ABBR_abb	ORG	ENTY_plant		LOC_state	GPE
ABBR_exp		ENTY_product	PRODUCT	NUM_code	PERCENT, CARDINAL
DESC_def		ENTY_religion	NORP	NUM_count	CARDINAL
DESC_desc		ENTY_sport	EVENT	NUM_date	DATE, TIME
DESC_manner		ENTY_substance		NUM_dist	QUANTITY, CARDINAL
DESC_reason		ENTY_symbol		NUM_money	MONEY
ENTY_animal		ENTY_techmeth		NUM_ord	ORDINAL
ENTY_body		ENTY_termeq	LANGUAGE	NUM_other	PERCENT, CARDINAL
ENTY_color		ENTY_veh	PRODUCT	NUM_perc	PERCENT
ENTY_cremat		ENTY_word	GPE	NUM_period	DATE
ENTY_currency	MONEY	HUM_desc	PERSON	NUM_speed	QUANTITY, CARDINAL
ENTY_dismed	ORG	HUM_gr	ORG	NUM_temp	CARDINAL
ENTY_event	EVENT	HUM_ind	PERSON	NUM_volsize	QUANTITY, CARDINA
ENTY_food	PRODUCT	HUM_title	WORK_OF_ART	NUM_weight	QUANTITY
ENTY_instru		LOC_city	GPE		
ENTY_lang	LANGUAGE	LOC_country	GPE		
ENTY_letter		LOC_mount	LOC		
ENTY_other		LOC_other	GPE, LOC		