COMP90042 Assignment 2 Project Report

nchang1 858604 Kaggle: nchang1_occ

1. Introduction

The task of the project is to develop an information-retrieval based question answering system, which, given the question, retrieves answers from a set of Wikipedia documents.

2. Main System Components – V1.0

The V1.0 system is built following the general components on Question Answering system [1][2].

2.1. TF-IDF Based Retrieval Module

Based on the question, the retrieval module searches through the given document for the most similar text span, which in the system is set on the sentence level.

Both the question and paragraph are fed into a pre-processing function, where the stopwords are removed, and the tokens are normalised.

Term frequencies and document frequencies are calculated next, and the most frequent terms are filtered out to decrease noise to the search result (threshold set as half the number of text spans).

The similarity is scored based on a modified BM25 model, where the parameters are adjusted to highly favour the rare terms, and ignore the query term frequencies.

A ranked index based on the input sentence list is returned from the retrieval module.

2.2. Question Classifier Module

Question classifier tags the question based on the expected answer type. In the basic model, the questions are classified according to hand-written rules based on the keywords appearance, and four tags are used – LOCATION, NUMBER, PERSON, and OTHER.

2.3. Named Entity Tagger Module

The target of this module is to tag the input text span according to its named entity. The same tagset from the question classifier is used here.

The input text will first be tagged using Standford's NER tagger [4]. However, due to the limitations of the tagger, the result needs to be further processed.

- The tokens that contain numeric characters are tagged as NUMBER.
- The tokens that are months or ordinal are tagged as NUMBER.
- The ORGANIZATION tag is grouped under OTHER.
- The neighbouring tokens with the same entity tag are grouped together as one single entity.

2.4. Answer Extractor Module

Given the question and the tagged text, the answer extract module returns the normalised answer for the question. The input entity list is ranked based on the criteria below:

- whether it has appeared in the question
- whether it is tagged with the expected answer type

The best ranked entity will be returned.

3. Error Analysis and System Improvement

Overall, three major types of errors are observed:

- Passage / Sentence Retrieval Error
- Entity Selection Error
- Question Classify Error

Based on the errors above, the system is adjusted accordingly.

3.1. TF-IDF Retrieval – V2.0

System 1.0 used a two-step approach to search for the paragraph. First, it uses a set of keywords retrieved from the query, which contains only words from the below POS tagset to avoid noise from the verbs and wh-words.

$['NN','NNS','NNP','NNPS','JJ','JJR','JJS','CD','FW']$

Then a full set of query keywords (only with stopwords removed) is used to select the best matching sentence from the returned paragraph.

During performance testing, it is observed that the two-step approach results in higher error rate, because if the result from the first step is incorrect, the final result will definitely be wrong. Also, the system is making a lot of empty guesses, due to the absence of relevant named entity.

System Version 2.0 directly searches for the best matching sentence on the document level. And a back-off mechanism is added to the system, to keep searching for the next best-matching sentence, if no matching entity is found in the previous search. This improved the performance roughly by 0.03.

3.2. Answer Extractor - V3.0

System V1.0 uses two parameters in the process of named entity selection, whether the candidate appears in the question, and whether it is the expected answer type.

However, in case of multiple candidates, it will randomly pick the first entity on the list, as indicated in below example.

Sentence	IBM used liquid cleaning agents in circuit board assembly operation for more than two decades, and six spills and leaks were recorded, including one leak in 1979 of 4,100 gallons from an underground tank.		
Question	How many gallons of liquid cleaning agent leaked from an IBM facility in 1979?		
Entity	(u'two', 'NUMBER'), (u'six', 'NUMBER'), (u'one', 'NUMBER'), (u'4,100', 'NUMBER')		
Answer	4,100 gallons	**Predict**	two

To address this issue, a new parameter is introduced to the selection process in case of multiple candidates:

- Entity's distance to the nearest open class word token that appears in the question.
- If the above parameter still returns multiple candidates, entity's distance to the nearest closed class word that appears in the question.

System V3.0 improved the performance to 0.1227.

3.3.Question Classifier – V4.0

The answer extraction is based on the expected answer type. But the rule-based classifier is very limited in the accuracy in prediction, as it relies on a very small set of keywords, and cannot predict questions as below. Through observations, it is noticed that this accounted for a large portion of errors.

Question	In what span did Universal produce westerns with Kirby Grant?		
Correct	NUMBER	**Predicted**	OTHER

Two supervised machine learning models are developed to attempt to predict the question types -- a neural network based prediction model and a Naïve Bayes model. Both models are trained and tested with 5000 annotated questions [3]. The performance is as below.

Model	**Accuracy**
Neural Network	75%
Naïve Bayes	51%

Therefore, the neural network model is selected in System 4.0. 5000 entries from training.json is tagged based on the answer, and used in training the neural network model.

This improved the system performance to 0.1670.

3.4.NER Tagset

It is noticed that, even with the adjustment in Answer Extractor, the system still tends to select the incorrect candidates from the same tag. This is particularly obvious in NUMBER tag.

Therefore, a new detailed tagset is used, which further divide the NUMBER tag into DATE (e.g. 21 June 1991), YEAR (e.g. 1991), MONEY (e.g. $1), PERCENT (e.g. 1%), and NUMBER. The question classifier and named entity tagger are adjusted accordingly.

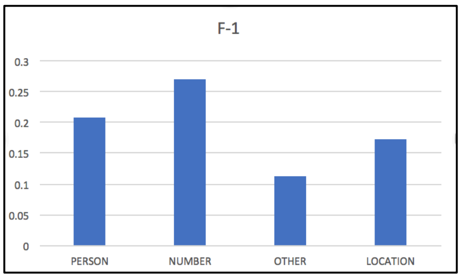
However, this modification set the system performance back by 0.05, which is therefore reverted.

3.5.System Performance Summary

F1 Score on Different Systems				
V1.0	**V2.0**	**V3.0**	**V4.0**	**Kaggle**
0.0715	0.1013	0.1227	0.1670	0.147

Different types of questions are evaluated separately, based on F1 score, precision and recall.

	F-1	Precision	Recall	Total Num
PERSON	0.2069437	0.22873995	0.19927614	373
NUMBER	0.27062169	0.29375661	0.26313492	756
OTHER	0.11236994	0.12386127	0.11127168	1730
LOCATION	0.17247899	0.18936975	0.17382353	238
Total	0.167	0.183	0.164	3097



4. Opportunities for Future Improvement

NER Tagger and Question Classifier

The current system uses Standford's 3-class-modelled NER tagger, which is trained mainly on recognising PERSON, ORGANIZATION, LOCATION tags [4]. This has resulted in certain accuracy loss, due to the discrepancy and generality of the tags. Increasing the class number on NER tagger and the question classifier has proved to be effective in previous study [2].

Context and Question trained Neural Network

With proper encoding method to incorporate question and context, Neural Network is also a high performing model in Question Answering [5].

References

[1] D.Jurafsky, J.H.Martin, Speech and Language Processing. 2017. Chapter 28.

[2] J.Lin, Exploration of Redundancy-Based Factoid QA. University of Maryland, 2007.

[3] Annotated Question Data, http://cogcomp.org/Data/QA/QC/

[4] Stanford University Named Entity Recognizer, https://nlp.stanford.edu/software/CRF-NER.html.

[5] B.Hicks, Neural Networks for Text-based Answer Extraction. Standford University, 2017.