Basic System - Reading Comprehension on Short Passages for Question Answering

JAKE PITKIN
CS 6390 - Information Extraction
March 10, 2017

Introduction

For my project I am going to explore a machine reading task. Given a short passage, the task is to answer a question, where the question is a span from the passage. Most work in the field of machine reading has focused on macro-reading: using massive corpora to answer factoid questions. This approach leverages the fact that the answer appears a redundant number of times. For my task, this isn't an option as the passages are typically a single paragraph. Instead this micro-reading task will need to leverage the structure of the dependency and constituency trees.

I will be using the SQuAD (Stanford Question Answering Dataset) (Rajpurkar et. al, 2016) dataset for this task. They provide 100,000+ labeled examples corresponding to 500+ articles. (*https://rajpurkar.github.io/SQuAD-explorer/*)

Approach

Due to time restraints, my basic system will be taking the approach of using lexical coverage. For each noun phrase in the passage, I determine how much word overlap the sentence the noun phrase appears in (ignoring the noun phrase) has with the question. The answer is chosen as the noun phrase with the most overlap. Character case is normalized and very frequent words are removed.

Over Spring Break, I'm going to overhaul my system and build a machine learning classifier which features that capture the structure of the passage using the dependency and constituency trees. This will allow me to fully leverage the large amount of annotated training data.

Technical Design

My project is written in Python 3 and uses the **Stanford CoreNLP** (*http://stanfordnlp.github.io/CoreNLP/*) and **NLTK** (*http://www.nltk.org/*) NLP toolsets. I use CoreNLP to predict dependency trees (to be used to my final system). NLTK is used to split the passages into sentences and tokenize each sentence as well as to find all the noun phrases in the passage.

The system design is organized as follows:

File Name	Description
`main.py`	program entry point
`ioutil.py`	I/O utilities
`nlptools.py`	API for various NLP tools
`evaluate-v1.1.py`	evaluates accuracy and F1 of predictions
`predict.sh`	script to run system and produce predictions
`evaluate.sh`	script to evaluate the predictions

Evaluation Results

I will be using 87,599 examples as training data (80%), 10,570 examples for development (10%) and 10,570 examples for testing (10%). Along with the dataset, an evaluation script was provided with the dataset I am using. It reports the percent of exact matches along with an F1 score.

I evaluated my basic system using the development data and compared to the results of making a random guess, a logistic regression model, and a human baseline provided by the paper accompanying the dataset (Rajpurkar et. al, 2016)

Approach	Dataset	Accuracy	F1-Score
My System	Dev	5.8%	10.7%
Random Guess	Dev	1.1%	4.1%
Logistic Regression	Dev	40%	51%
Human Baseline	Dev	80.3%	90.5%

These results make it clear that this task requires deeper understanding of the structure of the passages than my basic system provides. It demonstrates better performance than taking a random guess but is behind the ML model by a large margin. Other more subtle limitations include a preference towards noun phrases in longer sentences and the case when the answer

to the question is not a noun phrase. My system has a long way to go, but I have good ideas about the direction to take my development and I am optimistic about future results.

Appendix

Here I provide an example passage with an accompanying question and answer. The passages are separated into sentences and noun phrases are extracted (noun phrases are in bold). Finally, each sentence containing a candidate noun phrase is compared to the question to determine overlap.

Passage: When suffering from sleep deprivation, active immunizations may have a diminished effect and may result in lower antibody production, and a lower immune response, than would be noted in a well-rested individual. Additionally, proteins such as NFIL3, which have been shown to be closely intertwined with both T-cell differentiation and our circadian rhythms, can be affected through the disturbance of natural light and dark cycles through instances of sleep deprivation, shift work, etc. As a result, these disruptions can lead to an increase in chronic conditions such as heart disease, chronic pain, and asthma.

Question: What kind of deprivation results in diminished immune response and lower antibody production?

Answer: sleep deprivation

Sentence 1: 'When suffering from **sleep deprivation**, active **immunizations** may have **a diminished effect** and may result in lower **antibody production**, and a lower **immune response**, than would be noted in **a well-rested individual**.'

Sentence 2: 'Additionally, **proteins** such as **NFIL3**, which have been shown to be closely intertwined with both **T-cell differentiation** and our **circadian rhythms**, can be affected through **the disturbance** of **natural light** and dark **cycles** through **instances** of **sleep deprivation**, **shift work**, etc.'

Sentence 3: 'As **a result**, these **disruptions** can lead to **an increase** in chronic **conditions** such as **heart disease**, **chronic pain**, and **asthma**.'

Predicted answer: sleep deprivation

Overlap with question: 8 words

References

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.