Welcome to Gummi 0.6.6

January 14, 2019

1   Introduction

We are stepping into the world where machines will not just able to perform tasks by themselves but also be able to learn to improve their performance with the time. With this advancement in computing we can see a dream of transforming the human life will into something as we have seen or read in futuristic sci-fi fictions. We can argue that this dream might seem to be distant but initial footsteps of humans towards machine learning has shown great potential of what to unfold in future decades. Seeing this as golden opportunity, researchers and industry people all around the world have jumped deeply into it and now actively investing their time and energy to extract the realms of machine learning or modern deep learning. Because of so much interest going around the field, we are getting to see wonderful applications of machine learning popping up every other time.

As we have seen both the amount of data produced and hardware sophistication has taken a leap in last decade. With rise of Big Data technologies like hadoop, spark etc and hardware like GPUs, the machine learning practitioners are getting their hands dirty on different types of data and fully exploiting the strength of machine learning models. One such type of data is textual data. Internet alone holds the great chunk of world's textual data. It is been produced in the form of the blogs, social media, books question answering blogs etc. The practice of finding insights or patterns in the textual data is called text mining. One axis of text mining is Natural language processing (NLP). It is the branch of artificial intelligence which deals with the interpretation and generation of the human language. For example it takes human language generated text and provide applications like translation of text in multiple languages, find sentiments associated with the text or classification of text like in case to judge whether Email a span or not ?.

One of the fundamental task of the NLP is coreference resolution, the process of finding the expressions in the text referring to the same entities.

Lot of research has been done on this task dating back to 1970. Multiple models ranging from statistical to deep learning has been developed and thoroughly tested to create clusters like above efficiently. The reason for doing so much heavy duty is because coreference resolution has played a significant part in improving the NLP applications. More about about this in chapter 2. Seeing this prospect of it, I have decided to experiment it with the Question Answering (QA) system, a NLP application used for answering the user queries. I have used state of the art resolver and tested it on multiple data-sets. In remainder of introduction, I further motivate about the analysis and research I done in thesis and later I formulate the research questions to be answer in this study and finally the Section 1.2 will describe the outline of the thesis.

1.1 Motivation

Question Answering system is around from quite some time now and it is getting sophisticated over this time. In early times, it used to be closed domain system which answers certain type of queries but now we have moved from there to more open domain approach where multiple types of queries are handled. Examples of such modern QA systems are search engines like Google, Bing etc. We already know the importance these systems play in our daily life but we also understand it comes with the cost of digging in multiple documents to find conclusive answer. Researchers and NLP practioners are trying to make it more sophisticated to not only increase user experience but also spread true and clear information.

In this thesis three points have inspired me to analyse the effect of coreference resolution. 1. Is coreference resolution at all benifical for QA?. 2. If it so which type of resolvers will suit the best 3. Which subset of coreference relations e.g pronoun would be beneficial for the QA .

1.2 Thesis Outline

I have structured the thesis in four chapters. Chapter 1 defines the theoretical concepts. In Chapter 2 I defined the framework for incorporating the information in QA system. Chapter 3 is the evaluation of the results and finally the chapter 4 presents the framework for the user to test their coreference and finally I conclude the thesis with the points that can be further researched in future.