Cosine similarity as Machine Reading Technique

Gaurav Arora , Prasenjit Mazumder

Dhirubhai Ambani Institute of Information and Communication Technology
gaurav_arora@daiict.ac.in,p_majumder@daiict.ac.in

Abstract. Question answering for Machine reading evaluation track is a aim to check machine understanding ability of a machine.As we analyzed most crusial part for efficient working of this system is to select text which needs to be considered for understanding since understanding text would involve a lot of NLP processing. This paper covers our submitted system for QA4MRE campaign, Which mostly focuses on two part first being selecting text from comprehension and background knowledge needed to be understand and second being eliminating or ranking options based on selected text from former step.Our main focus was on eliminating and ranking which boils down to tunning various parameter for selection whether to answer particular question if answered how to consider scores,Following methods like calculating cosine between question and passage sentences,cosine of named entities output of passage sentences and question were also considered for scoring .In addition to this basic frame work of our system negation of sentences were also considered to answers which received very close score.We also considered expansion of question and options respectively to collect relevant information from background collection.Entity Co-referencing and normalization were some of important preprocessing to consider on passage and background collection as we analyzed since it can increase score of sentence or option which do not directly mention entity.

Keywords:
QA4MRE,openephyra,machine reading

1 Introduction

The irlab ,DA-IICT participated in QA4MRE campaing in CLEF 2011 to test how elimination strategry works in Question Answering in comprehension passage.Organized task tries to address basic problem of understanding text with comprehension passage questions.With a lot of breakthroughs in Natural Language Processing and Machine learning,text understanding have attracted a lot of researcher.Task aims to understand certain relevant portion of text from passage or background knowleadge to answer question with deep understanding of text.Major challenges to be addressed are selecting portions of text to understand as whole background knowledge and passage cannot be understood due to involovement of heavy NLP processing and real time response time constrain to

answer question.Second major challenge is selecting or eliminating option based on undertanding of understanding of choosen part of passage and passage.

Our main focus was on selecting relevant text to understand as further system depends on performance of this step.various techniques like cosine value of plain text,cosine value of Named Entity Tagged sentences ,combination of both.Various other variation like theasurus conversion befor finding cosine,Adding option to question were done to increase quality of relevant text extracted for understanding.One of problem while answering question is to select wheather to answer question or not .Selecting threshold for deciding to answers question was also a challenge to address.Since in question answering system its better to leave question if system feels it dont have enough understanding to answer underlined question.Broadly these were challenges we addressed participating in QA4MRE campaign

The paper is organized as follows. We describe the devel-oped system in Section 2, report the experiment results for 120 questions on 3 different topics in Section 3, and conclude in Section 4.

2 System Description

We describe our system for answering comprehension type question based on knowledge acquired from passage , background knowledge documents.To ansswer question system based on question first try to identfy text which should be understood,based on this identified text question is answered by elimination.Since for all question it is difficult to get proper text hence selecting wether to answer question is very crusial and is being addressed by selecting and checking various threshold values.

2.1 Text Selection for understanding:

To answer question sentences in comprehension passage need to be ranked based on relevance to the question .The dataset used for the track contains background collection on three topics and series of test containing single document test document with several question and set of choices per question provided by organizer of tracks.Since by manual inspection it was observed that most of answers were residing in passage,Finding out relevent sections of passage to questions was crucial.To address this issue cosine similarity sentence ranking was used. Cosine Similarity sentences Ranking: Cosine similarity I.e dot product of sentence and question was taken and dot product of both were taken as similarity measure for ranking sentence and as due to absence of terms in question in comprehension passage many times passages sentences were not ranked appropirately I.e any junk sentences were getting a higher rank.So as a conter measure to that we found solution to be add option which is provided to us in question while we are maching it with sentences of comprehension passage This actually helped us in increasing scores of good sentences for question.by altering this relevant sentences were ranked higher . This ranking of sentence gave score to each of sentences

in passage i.e relevance score w.r.t question to each sentence.Score was used to judge wheather sentence should be considered for answering or not.manual inspection was used to decide upon the threshold of sentence selection. Since we mentioned sentence selection is very critical for working of overall system,Cosine Similarity sentence ranking performed wasnot able to get any relevent sentences for a lot of questions.Few methods like Theasurus conversion and option addition were tried to improve the accuracy of sentence.techiniques are descibed below

Theasurus conversion: Since all resources options , questions and best ranked sentences of passages can have different style of words and written with different language so it is necessary to convert all important words to same domain of same word.. Option addition: In cosine similarity sentence ranking method we found cosine similarity with question in this we added all options to the question string while finding cosine similarity.This helped us in finding relevent sentences for a lot of sentences were cosine similarity method failed to score relevent sentences higher,in some case it also incresed false positive rate i.e selection of unrelevent sentence .

Other than cosine similarity we tried cosine similarity of Named Entity cosine similarity,in this method cosine similarity was calculated for Named entity tags of sentences in passage and question to score sentences with relevance.but this method was dropped later on as it was not ranking sentences properly as sentences had very few named entities and ranking just based on type of tags present in it leads us to very less named entities which made scoring irrelevant.

Atlast Cosine similarity sentence ranking with Theasurus conversion and option addition was used to select relevant sentences for ranking and text understanding which in performed upto 85% on sample data provided by track organizer.after this we got few selected sentences from passage relevent to answer question.Next major challenge is Ranking option.

2.2 Ranking options:

With availability of relevent sentences for question next task is based oon nderstanding of given text answer ,select answer or eliminate options.Since there are 5 options to each question and answering question in this format boils down to just ranking options based on some criteria and choosing best options as answers.Option ranking was done considering calculated score from Cosine value of option measured with best couple of sentences ,Semantic Value of options,comparing options with information from collections is also considered.Cosine Similarity option ranking ,semantic value etc are descired below.

Cosine Similarity option Ranking: Cosine similarity i.e dot product of best sentence out of passage which are selected by emperically selecting threshold to see best sentences are only choosen and options was taken.then these score of sentences are processed and updated through various modules like Semantic Meaning etc.

Order Sequencing of Named Entity: Order in which named entity apprear is very crucial as they make meanng of sentences so option + question and best

sentences are matched for ordering of Named Entities.Entities if are in order then relevance of option is very higher.

Semantic Meaning of Options and Questions : Senti Wordnet was used to find out the essence of sentences .It is generally observed that answer to questions are genrally have positive sentement value so to check that if the value of options is negative then the sentence is ranked lower and otherwise cosine value is taken as score of options.

Using Above Mentioned techniques the Scores were assigned to verious sentences/options and best sentences i.e sentences with highest cosine value were choosen as relevant and displayed as answer.

Above technique resulted in low ranking score for almost 35 % of questions so for those question we tried cosine similarity with retreived passages indri framework was used to retreive pasaages from the background collection.then similarity measure is calculated with reterived relevent passages from background collection.length of reterived passage was fixed to be 200 characters based on paper by Tellex, Stefanie,MIT Artificial Intelligence Laboratory.

2.3 Answering Question

As our system is based on ranking of option even if no text or relevant passage is extracted by system then to some neglegible score will be assigned to the the option and while answering we need to identiify such question and leave these kind of question unanswered.With manual inspection we found out some 15% of questions were getting very low score so to avoid answering such questions we submitted runs to evaluate the threshold and if sum of all options is less than 0.1 then question was left unanswered.

References

1. OpenEphyra System :- OpenDomain Question Answering System. http://www.ephyra.info/
2. QA4MRE Final Guildline :- TECHNICAL COORDINATION, Pamela Forner, CELCT, Italy
3. Indri 3.0 :- opensource project maintained under apache licence.
4. Answering and Questioning for Machine Reading Lucy Vanderwende,Microsoft Research ,Redmond, WA 98052
5. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and