Evaluation of Extractive Summarization Techniques on Powerpoint presentations and HTML pages from the AAN TutorialBank Corpus

Yijiao He, Alexander R Fabbri, Dragomir Radev
Department of Computer Science, Yale School of Engineering and Applied Sciences, New Haven, CT

LILY Lab

Introduction

As the body of knowledge and research increases around AI and NLP, there is an increasing need for a tool that can be used to summarize the key points of knowledge areas at a higher level, such that beginners in a field are able to quickly grasp the basics without getting mired in the details of a topic. This project aims to compare the performance of different extractive summarization algorithms on Powerpoint presentations and html webpages, leading towards the creation a tool that takes in a scientific topic as input, and can produce a coherent set of factoids relating to that scientific topic. More specifically, the project will use the topic that one desires to obtain a summary for as a topic keyword, and perform information retrieval on the AAN dataset of presentations and webpages of topics relating to NLP. Most previous attempts have focused on the summarization of scientific articles, whereas Powerpoint tutorials and HTML webpages are generally less structured. Thus, we investigate methods on extracting salient information from these types of resources to automatically generate a survey of the subject.

Materials and Methods

The documents used were from the AAN TutorialBank Corpus. 200 topics were chosen to perform annotations on. For each topic, 5 relevant resources were chosen and annotated for relevance to the topic. The summarization was done in 2 steps, using 7 query topics that both had student annotations of slides and manual summaries. The first step was information retrieval. All resources were run through a ranked information retrieval algorithm and scored for their relevance to the given topic. The top 10 resources were then picked for step two. The second step was summarization. The 10 relevant resources were fed through four different extractive summarization algorithms. The manually generated summaries from Jha et al. were used as gold standard summaries to evaluate against the results. The evaluation was performed using BLEU, and a topic extraction program.

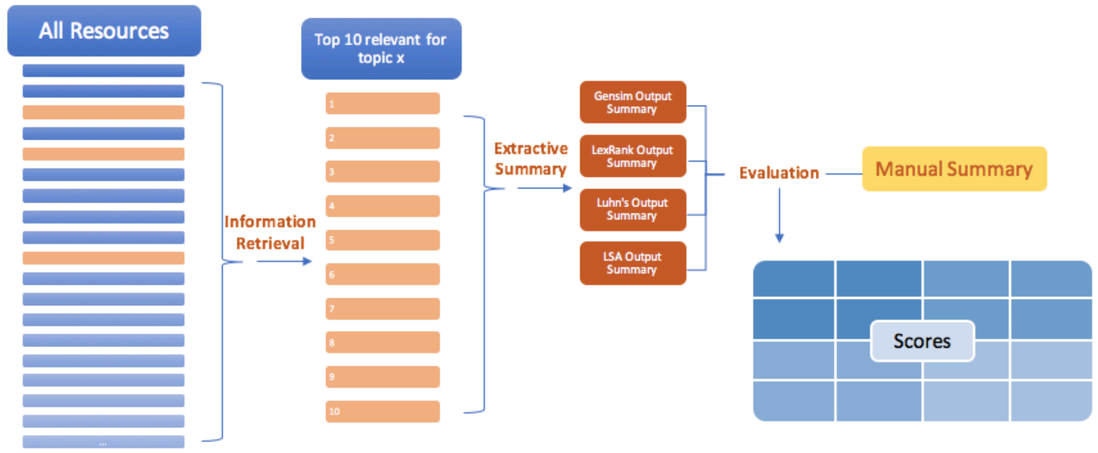

Figure 1. Extractive Summarization process.

Table 1. mean BLEU scores on summaries from different data categories after information retrieval.

Algorithm	Combined	Individual Slides	HTMLs
Gensim	0.56	0.13	0.63
Lexrank	0.68	0.09	0.70
Luhn's	0.48	0.19	0.46
LSA	0.57	0.21	0.50

Table 2. Topic Extraction results from different data categories after summarization.

Algorithm	Combined	Individual Slides	HTMLs
Gensim	-1	6	0.5
Lexrank	-2.5	5.5	0.5
Luhn's	1.5	5.5	1
LSA	-1.5	4	-2

Table 3. Excerpts of sample outputs from various summarization algorithms on the topic of "Word Sense Disambiguation".

Resource Type	Excerpts of output summaries
Luhn on HTML	These annotated documents are used to evaluate our word sense disambiguation systems in "Semi-supervised Word Sense Disambiguation with Neural Models", Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans and Eric Altendorf, Proceedings COLING 2016##Word sense mappings This package also includes maps from NOAD word senses to WordNet senses.
LSA on Individual Powerpoints	Word Sense Disambiguation SENSE DISAMBIGUATION the ability to computationally determine which sense (word/phrase) is activated by its use in a particular context of 52 144 Semantic Similarity between word senses Multilingual Joint Word Sense Disambiguation(MultiJEDI)
Manual Summary	Given a polysemous word in running text, the task of WSD involves examining contextual information to determine the intended sense from a set of predetermined candidates. Word sense disambiguation (WSD) has been found useful in many natural language processing (NLP) applications, including information retrieval

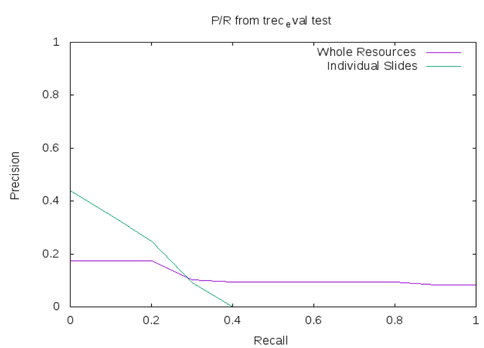

Figure 2. Precision Recall output from information retrieval using Trec_Eval.

Results

The information retrieval portion of the summarizer achieved a mean average precision of 0.08 and 0.10 for individual Powerpoint slides and combined resources respectively.

The BLEU scores indicate that HTMLs performed better across the board compared to individual slides. This is likely because sentences from presentation slides tend to be shorter and more succinct, which do not have the variety of words when compared to HTMLs. By summarization algorithm, Lexrank seems to be the best performer for combined resources as well as HTMLs, but performed most poorly on individual slides. This could be because Lexrank's avoidance of repetition in the final output may have eliminated certain sentences coming from slides, which could further bring down the BLEU score. LSA performed best with individual slides and had average to above average performance for both combined resources and HMTL webpages.

For Topic Extraction, Luhn's algorithm had the best performance on the summaries generated from the information retrieval, while LSA performed the worst.

Conclusion

While the BLEU results for individual slides in this exploration did not yield better results than the other formats, this is still promising in that the summaries were legible and still yielded relevant factoids. Especially when considering topic extraction, in comparison to the other resource formats, the individual slides' higher likelihood of extracting the relevant topic indicates a higher direct relevance to the topic.

It is worth investing time in the future to write a better parser that can ignore slide titles and page numbers, as well as webpage reference links and disclaimers.

BLEU is not necessarily the best tool for measuring summarization results, as it was intended for machine translation, thus another measure such as using Pyramid Scores might give results that are more meaningful for our purpose. Which is also the future direction of this work.