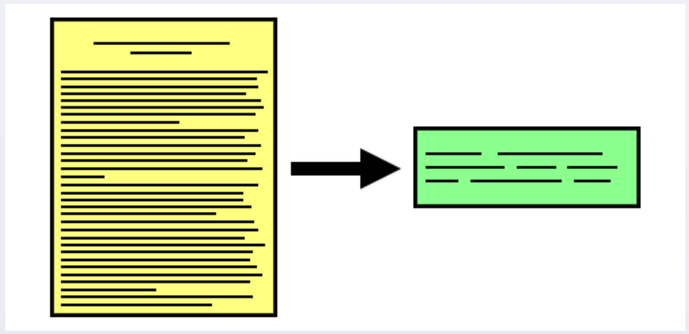
AUTOMATIC KEYPHRASE EXTRACTION

Avinash Mohak, Mentors: Arijit Biswas, Ankit Gandhi

Xerox Research Center India

INTRODUCTION

Key phrases provide a concise and meaningful summary of a document. They are extensively used in document indexing, categorization, clustering, search and summarization. Despite its importance, state-of-the-art performance in automatic keyphrase extraction is still much lower than many core NLP tasks. Several factors contribute to this difficulty, including document length, structural inconsistency, changes in topic, and (a lack of) correlations between topics. Moreover, there is no absolute ground truth for the keyphrases, except human-labelled keyphrases, which varies across humans!



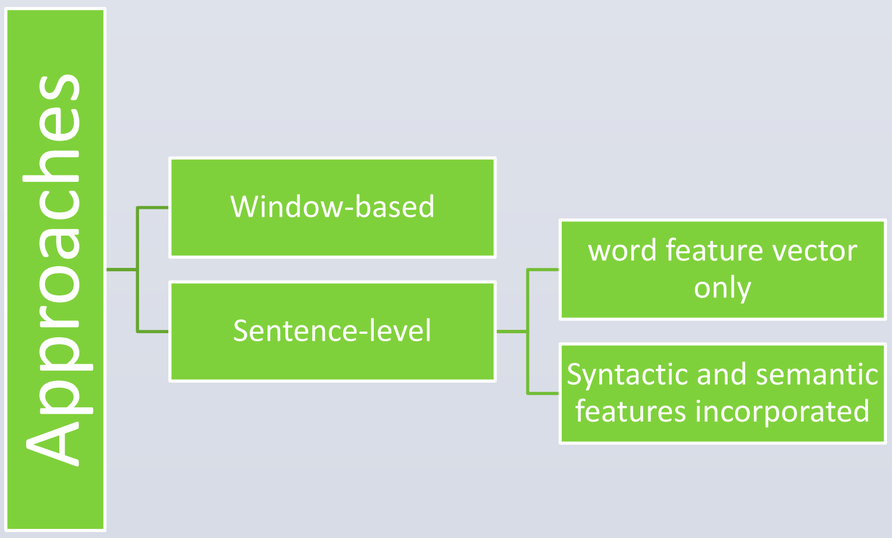
OBJECTIVES

We aim to identify the key phrases in any given document. Particularly, we focus on the subtitles of the educational videos(lectures) in the field of Computer Science. We simplify our goal by identifying keywords in the document and leverage its results in identifying the key phrases.

MOTIVATION

Our approach relies on the basic idea: Keywords are a function of their context. In other words, it's the context or its adjacent or nearby words, which determine "keywordness" of any word. Since recurrent neural network (RNN) are particularly designed for capturing the context, we used it as our underlying model, expecting it to learn to identify the keywords in any given text. Specifically, we used LSTM-based model, which does off with the exploding/vanishing gradient problem (which exists in vanilla RNN model), thus allowing the model to capture a longer context and, thus, learning in a better fashion.

APPROACHES



ARCHITECTURE

Input Window – Takes in a sequence of words (indices) and passes it to the next layer

Look Up Table Layer - Converts the index into the specified dimensional real-valued vector and learns a relevant representation for each word, trained by backpropagation

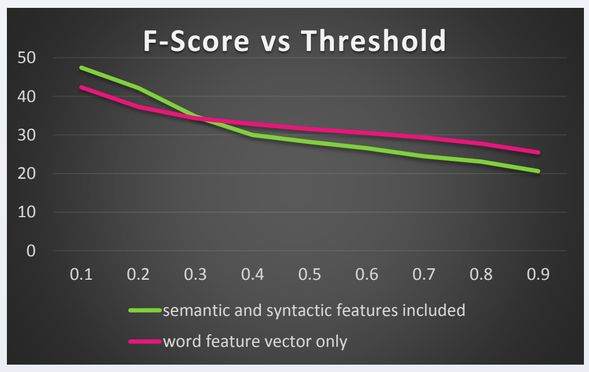
LSTM Cell Layer – Takes a sequence of input vectors and outputs a sequences of target vectors

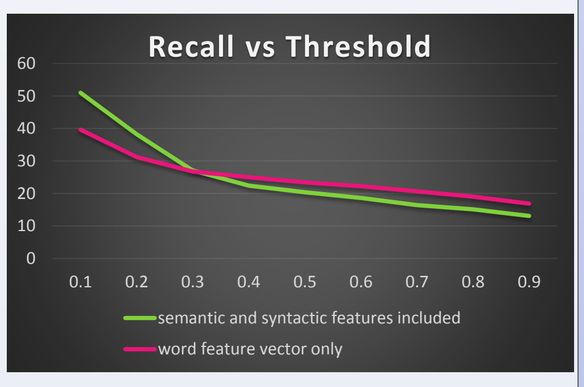
Linear Layer – Standard NN layer which applies a linear transformation to the input vector

LogSoftMax Layer - Applies the function to an n-dimensional input Tensor

RESULTS

We present some of the testing results which compares the two sentence-based approaches, which cue us to use the optimum approach and threshold for the application phase.





As a final step, we tested the model on the text we aimed at. Though the performance was reasonably well. Due to lack of any performance measures for it, we present few good and not-so-good examples of our model as applied on the subtitles of the lecture.

Good:

1. **Optical signals** traveling for long distances through **fiber** need to be strengthened.
2. **Multiplexing** is about sharing a medium that means different users are sharing the same medium for communication at the same time.
3. As I mentioned, you have this **geostationary satellites**, which are nothing but repeaters on the sky.

Not-So-Good:

1. One issue which is important in MAN is the issue of **access**.
2. We have seen the different ways these **digital signals** and analog signals etc. can be used for communication and how digital **data** or analog **data** can be encoded.
3. you can use **multimode** fibers over here and you can use single mode fibers

CONCLUSIONS

- Keywords can be modeled to be a function of their context quite reasonably.
- Word representation (vector embedding) affects the performance of the model.
- Syntactic and semantic features of a word play a considerable role in determining its importance.

FURTHER WORK

Our work can be easily extended to a paragraph-level approach where we input an entire paragraph to the model, and then predict the keywords. Besides, we can leverage the results from classical approaches viz. graph-based ranking algorithm or topic-based clustering, in an ad hoc or some heuristic method.

REFERENCES

1. Collobert, Ronan, et al. "Natural language processing (almost) from scratch."Journal of Machine Learning Research 12.Aug (2011): 2493-2537.
2. Hasan, Kazi Saidul, and Vincent Ng. "Automatic Keyphrase Extraction: A Survey of the State of the Art." ACL (1). 2014.
3. Hasan, Kazi Saidul, and Vincent Ng. "Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.
4. Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).
5. Nallapati, R., Zhou, B., glar Gulçehre, Ç., & Xiang, B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond.
6. Lipton, Zachary C., John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning." arXiv preprint arXiv:1506.00019 (2015).

ACKNOWLEDGEMENTS

We acknowledge Xerox Research Center India (XRCI) for supporting and funding the research work. Special thanks to Om Deshmukh for providing us the opportunity to work together on such an interesting problem.