Keyword Extraction by Deep Learning
Mid Term Report

Yi Cheng(yicheng1), Anoop Hallur(ahallur), Xiaoqiu Huang(xiaoqiuh)

October 31, 2014

1 Introduction

As the goal of our project, we plan to do keyword extraction using deep learning. We are aiming to improve the performance of keyword extraction by deep learning techniques as compared to other techniques used prsesntly

2 Progress

As part of our schedule, we had planned to

1. Be familiar with existing techniques and related works.
2. Learn *Deep Learning* methods and how it can be used in NLP related applications.

by the mid term schedule

To evaluate the performance of our approach compared to existing techniques, we have used Precision and Recall rates as our benchmarking standards.

As part of (1), we have implemnted TF-IDF (and Lexical Tree approach) for Keyword Extraction. With plain vanilla TF-IDF, the precision and recall rate is less than 10 %. However, after preprocessing the data by using a stemmer, the accuracy is a modest 15-20% on topics in our data set. We have used portland stemmer, available as part of open source NLTK(Natural Language Processing Toolkit) project.As part of survey paper[6], we found out that the existing standard for key word extraction uses Lexical Tree approach and we are coming up with the precision and recall rates for Lexical Trees approach
{Report exact figures here}.

As part of (2), we have prepared the following flow for the midterm report.
{Insert that block diagram here}.
{Insert more details here later}

3 Dataset

We had identified data from stackexchange(`https://archive.org/download/stackexchange/stackexchange_archive.torrent`) to be a suitable data set for our project. Of the 20 GB, available to us, we are using only 500 MB of data, on about 20 topics for our evaluation. We felt this was enough for testing purposes, and once we decide on the model, we will make use of all the data set available to us. Working with bigger datasets is a hassle when no concrete model is available and hence we stopped at 500MB. All our results and observations have been made on this 500MB of dataset, which we have cleaned and formatted for our applications.

4 References

1. Bengio Y, Schwenk H, Sencal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

2. Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013: 1631-1642.

3. Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011: 129-136.

4. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

5. Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information[J]. International Journal on Artificial Intelligence Tools, 2004, 13(01): 157-169.

6. Lott B. Survey of Keyword Extraction Techniques[J]. UNM Education, 2012.

7. Hasan K S, Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art[J].