Keyword Extraction by Deep Learning

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

September 29, 2014

1 Introduction

Recently, *Deep Learning* has been successfully applied in some tasks of *NLP*, such as parser, machine translation, text summarization and so on. Keywords extraction is a very important subfield of *NLP*, which is very useful especially in *Information Retrieval*, but little work have been done with using the method of *Deep Learning*.
In this project, we plan to implement keyword extraction of short and long textual data by application of deep learning. We would like it learn the highest ranked words in it by constructing a suitable models of the words in the document.

2 Dataset

We have identified data from stackexchange(`https://archive.org/download/stackexchange/stackexchange_archive.torrent`) to be a suitable data set for our project. It has labelled data from various topics such as programming, travel, philosophy etc. We have apporximately 20 GB of labelled data. This is large data set to test our algorithms on a larger scale. For our project we plan to train on a specified portion of the data and test on the remaining data. Since the data is cleanly labelled, we need not do any special processing/cleaning up of the data set other than splitting it into smaller segements for it to be suitable to work on.

3 Plan and Milestone

Since in this project, we try to apply *Deep Learning* to *Keyword Extraction*, we divide the project into the following small tasks:

1. Be familiar with existing techniques and related works.
2. Learn *Deep Learning* method and how it can be used in NLP related applications.
3. Design our model and experiment with the parameters and try to improve the model.

4. Analyze the performance on the test data set and summarize the results in the project report.

For the midterm report, we aim to finish the first two tasks, i.e implement some of the previously proposed methods and test them on the stackexchange dataset.
Here is a list of papers need to be read before the midterm report:

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

- 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.

- 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.

- 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.

- 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.

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

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

4 Goal

We need to apply *Deep Learning* to the problem and do experiments on it. If the performance is worse than that of the previous works, **at least** we need to figure the reason as to why the performance is poor and how to improve it.
As a stretch goal, we need to create the model which can automatically generate new keywords and finally those new keywords can be combined with our model to form a novel semi-supervised word extraction algorithm.