University of Sussex

My Theory of Everything

How it all works

Joe Bloggs

Submitted for the degree of Doctor of Philosophy

University of Sussex

September 2008

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Joe Bloggs

UNIVERSITY OF SUSSEX

Joe Bloggs, Doctor of Philosophy

My Theory of Everything
How it all works

Summary

Acknowledgements

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Word embeddings is a method of word representation mainly used for natural language processing. It consists of vocabularies and its positional information in euclidean space. This positional information then can be used to infer interconnection between words, whether its similarities or usage of the word in a sentence{CITE}. To obtain word embeddings, a model is created to extract features of a word from a corpus and mapped its location in euclidean space based on the features found. These word embeddings then can be used to do many downstream tasks, such as postagging, machine translation, and sentiment analysis. In general, large corpus with many words and examples of word usage is prefered because the size of the vocabularies will increases and more words connection can be inferred{CITE}. However, it is not posible to have such corpus since the language itself is evolving and there are many cases of typography in a documents and these words maybe not present in the corpus hence not included in the vocabulary. This words are called as *out-of-vocabulary* (OOV) words. Due to this problems, simple approach by asigning unique random embedding or by replacing OOV with an unknown $<UNK>$ token{CITE} . While in some cases using these simple approaches can produce acceptable results {CITE}, further improvement on downstream tasks can be achieved by using machine learning method to infer OOV embeddings.

To infer OOV embeddings, the model is built over quasi-generative perspective. Only knowing the vocabularies and its embedding, the model tried to generate embedding for OOV words, the results are tested for downstream task namely postagging and compared to previous state of the art to search for improvement over previous models.

Chapter 2

Related Works

Word2vec is one of word embedding that is trained using skip-gram model {CITE}. A word $w(t)$ used as an input and its context word, for example context word with windows of 4 $w(t-2), w(t-1), w(t+1)$, and $w(t+2)$, used as the target and the projection from input to the output is used as the embedding of the input $w(t)$. This model is highly dependant on the corpus completeness. More examples and vocabulary a corpus has the better the representation of the embeddings.

postagging explanation

mimick explanation

cnn-grams approach explanation

Chapter 3

Conclusion

I was right all along.

3.1 What was I right about?

I was right about the following things.

3.1.1 Previous theories were wrong

People thought they understood, but they didn't.

3.1.2 My new idea is right

Of course.

Bibliography

Appendix A

Code

```
10 PRINT "HELLO WORLD"
```