Semantic Embeddings in IR

Text-Based Information Retrieval

Joren Verspeelt, Jochem Geussens, Vincent Tanghe

April 6, 2016

1 Introduction

In the field of Natural Language Processing (NLP) and Information Retrieval (IR), there is a large need for good presentations of text documents. Recently, a lot of researchers in the field are presenting dense word representations (also known as Word Embeddings, Neural Embeddings or Sementic Embeddings) as a good solution. For the Course of Text-Based Information Retrieval, we are given the opportunity to work with state-of-the-art algorithms. The goal of this assignment is to gain practical experience with these algorithms by comparing them and using them in an application. First we will test and discuss different techniques on a [as a warm-up, blabla – got tired of writing]. Then, we will discuss our implementation for a search engine to find pictures based on their description.

Contents

2 Part I (Warm-up)

2.1 Task description

In this part, we will implement a small analogy solver that tries to guess the correct word, given simple analogy questions. E.g.:

$$a : b = c :?$$
$$Bratislava : Slovakia = Bishkek :?$$
$$ate : eat = found :?$$

Then, we will run several analogy solving models with several different representations on the benchmarking analogy dataset.

2.2 Setup

2.3 Results

2.4 Discussion

3 Part II

3.1 Task description

3.2 Setup

3.3 Results

3.4 Discussion

4 Conclusion

[Strong points, Weak points, lessons learned]

5 References