Super awesome embeddings

Grzegorz Beringer, Mateusz Jabłoński, Piotr Januszewski, and Julian Szymański

Faculty of Electronic Telecommunications and Informatics
Gdańsk University of Technology, Gdańsk, Poland

Abstract. Abstract

Keywords: word sense disambiguation, word embeddings

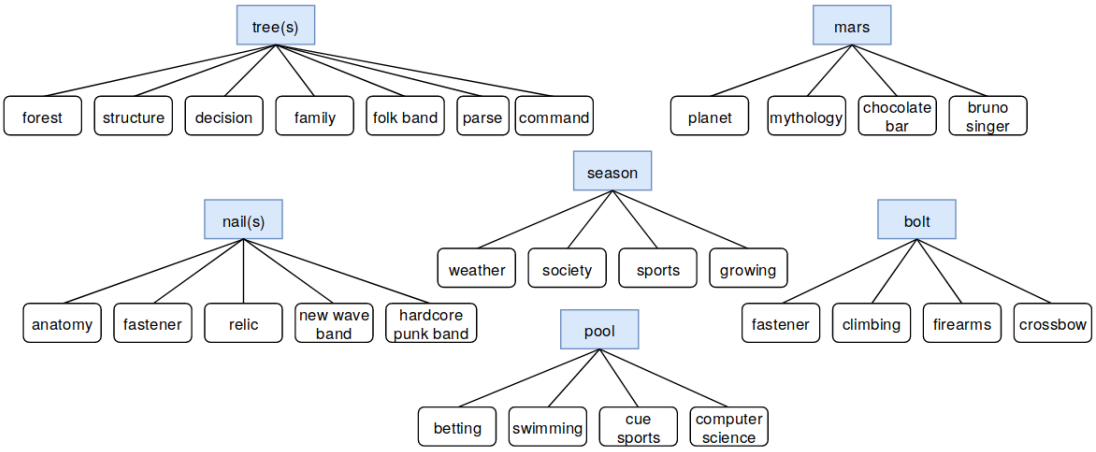
1 Introduction

2 Related work

3 Dataset

For the purpose of testing word embeddings as a method to differentiate between different meanings, we gathered examples for 6 ambiguous words, 4-7 meanings each (28 meanings in total). Ambiguous word together with its meaning constitues a *keyword*, which we use as a separate class when identyfing the closest meaning given some context. All keywords can be seen on Figure 1.

Fig. 1. Ambiguous words with their meanings (keywords) from the dataset



Examples were mostly gathered from Wikipedia, using *What links here* utility for each keyword. If usage examples from Wikipedia were not enough, other websites were used (or even the Wiki article on specific keyword itself).

The dataset is split into training and test set, with 5 training and 10 test examples for each keyword. Each example is stored in plain text, with the ambiguous word marked with "*" on both sides.

Example for *bolt crossbow* keyword:
*"Sulfur- and oil-soaked materials were sometimes ignited and thrown at the enemy, or attached to spears, arrow and *bolts* and fired by hand or machine. Some siege techniques—such as mining and boring—relied on combustibles and fire to complete the collapse of walls and structures."*

The correct keyword for each example, together with a path to file and link, where the original text was taken from, are stored in CSV files: *train.csv* for training set, *test.csv* for test set. Keywords themselves, together with links to their Wiki articles, are stored in *keywords.csv* file.