| Man | Women | King | Queen |
|---|---|---|---|
| Athens | Greece | Oslo | Norway |
| great | greater | tough | tougher |

Table 1: Word analogies used in word embedding validation

# 1 Background

*Embedding*
Machine learning (ML) tasks rely on a numerical (vectorial) representation of text which we refer to as an embedding. These can be calculated for texts of different lengths such as a title, sentence, paragraph or an entire document[**?** ]. Word embeddings are these numerical representations of a word, these vectors are an distributed representation of the word over the multiple (vector) dimensions(**?** ]). The word embeddings can be used to construct embedding of larger texts. Word embeddings can capture both the semantic and syntactic information of words. The advantage of the machine learning models is that it can be done without human-interaction(**?** ]). Word embeddings have improved various Natural Language Processing (NLP) areas such as named entity recognition, part-of-speech tagging, parsing, and semantic role labelling (**?** ]).

*Embedding validation*
Embedding validation techniques are methods that are used to validate the quality[1] of an embedding for a specific task(**?** ]). Multiple validations of word embeddings have been used, including: Word analogy, text similarity, categorization and positional visualization.

Word Analogy
Word analogy validation is based on a labelled validation set, containing, commonly, word pairs of four, that can be logically divided into two parts. As Table 1 shows, each last word can be derived from the three words before. The score is the fraction of correctly given fourth words, given the first three words. This validation metric is used in multiple studies[**?** **?** **?** **?** ]. Both this validation technique and the Word Similarity technique use vector distance calculations to validate the embeddings, this can therefore also be written as:

$$X_{Man} - X_{King} \approx X_{Women} - X_{Queen}$$

This means that the resulting vector of embedding of "Man" minus the embedding of "King" is approximately the embedding of "Woman" minus the embedding of "Queen". This resulting vector may be close to a vector "monarch" for example.

Word Similarity
A method to test the quality of word embeddings is the word similarity test. For these test, the distance between the word embeddings (vectors) is measured and compared to similarity scores defined by humans. Multiple non domain specific validation sets are publicly available including: the Rare-word dataset introduced in the paper "Better Word Representations with Recursive Neural Networks for Morphology" by **?** ], the MEN test collection by **?** ] and the WordSimilarity-353 test collection by **?** ]. These sets, among others, have been used in multiple studies of word embeddings[**?** **?** ]. This validation metric also relies on labelled data.

Classification
The classification validation method is a simple task which compares multiple texts. **?** ] used data from StackExchange and tried to determine if a pair was a duplicate. Even though the validation method is simple, it too used labelled data to validate the acquired results.

Categorization
**?** ] used for their research a dataset of IMDB with 100,000 movie review. They validated their proposed paragraph vector model by determining whether a review was positive or negative.

Position Visualization
(Unlabelled, Needs human validation)**?** ] and **?** ] mapped their word embeddings to a two

[1] With quality we mean the extend to which the task is completed correctly

dimensional vector to be able to display them in a graph and applied colors to various categories. The advantage of this is that a human can directly see that the embeddings make sense, however this approach is not applicable by a computer.

Even though these validation methods are not limited to domains, the labelled data they use are, since they do not consist of words of specific domains. At this moment, there are no sets for every domain which make it difficult to compare the accuracy of domain specific word embeddings to non domain specific word embeddings.