Motivation

Information retrieval
Information Retrieval(IR)is the activity of gathering relevant information, given another initial piece of information. The most practical example of this is a search engine. Given one or more search words the search engine will attempt to find relevant information. For example, a Google search on "Information Retrieval" (initial piece of information) will give you a list of results (relevant information). To be able to do this, the computer (search engine) must know which texts are related. To achieve this, multiple techniques can be used, such as TF-IDF, Word2Vec, Paragraph Vectors and GloVe. These techiques can be divided into two categories, ones that use a neural network and ones that do not.

Neural Network and Information Retrieval techniques
Neural Networks are computational structures, based on vector and/or matrix calculations. They can be applied to many different tasks, but need to be trained for each task. This training is the process of iteratively adjusting multiplication matrices or vectors to achieve the optimal result. Which is the result that is as close as possible to the given output for different input and output sets. In some IR techniques, Neural Networks are used to either predict words that may occur around a given word, or predict a word given words that surround it. For example, given the sentence

"The search engine searches for relevant information"

the Neural Network can be trained to either predict the words around "searches" (the, search, engine, for, relevant, information) or to predict the word "searches", given the surrounding (the, search, engine, for, relevant, information). This trained matrix (vector per word) now indicates "word relatedness" which, as mentioned earlier enables association (thus retrieval) with related(relevant) texts.

Word2Vec
Word2vec word embeddings are created using neural network, Word2vec learns word embeddings via maximizing the log conditional probability of the word given the context word(s) occurring within a fixed-sized window. Therefore the learnt embeddings contain useful knowledge about word co-occurrence[**?**]. There are multiple input/output possibilities for the neural network, best known are Skip-gram and the Continuous Bag-of-Words model (CBOW). The Skip-gram model takes a target word as input and outputs the predicted output words, while CBOW takes the context words as input and outputs the predicted target word[**? ?**].

Paragraph vectors
Variations on the word2vec model have also been proposed, **?**] introduced the paragraph vector, based on the word2vec model. The paragraph vector model uses additional variables to improve the accuracy of the word-embeddings. An advantage of the paragraph vector model is that it takes the word order into consideration, atleast in a small context [**?**].

GloVe
?] introduced the GloVe (Global Vectors) model. This model captures the global corpus statistics. The model transforms the word co-occurrences of all words in the corpus to chances, it excludes all the zero values and uses that as initial input for the neural network.

TF-IDF
TF-IDF is an abbreviation for Term Frequency times Inverted Document Frequency. This method does not rely on a neural network and therefore does not require training. The TF-IDF score is the product of the term frequency in a text and the inverted document frequency of the same term in a corpus of texts. Both of which can be calculated in a variety of ways.

Validation methods
The results produced by the techniques have to be validated to determine their quality (in usage). The quality of the results can be validated through, among others the F1 score and, for classification tasks, the rank of the correct category. The embeddings can furthermore be validated through their performance on tasks such as word analogies, word similarities, categorization and position visualization. These tasks can be designed to produce a score that gives an indication of the performance on a specific task. **?**] found that

a single validation metric cannot produce a representative result for other tasks. Embeddings that perform well on one task do not have to perform well on another task. As a result, the findings about performance of an embedding method are limited to the task on which they are tested, their results cannot be generalized to state that the embedding are overall "performing well". Validation tasks use either labelled or unlabelled data. Labelled data is data that is in some way marked, so that the correct answer can be derived from it, in contrast to unlabelled data.

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 **??** 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 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.

Domain specific
Earlier work on this topic, concerning domain specific articles, by (**?**]), found that in-domain training of the word embeddings can improve the process of document clustering. This effect is even stronger than the number of training examples and the model architecture.[**?**]. **?**] found that the corpus domain is more important than the corpus size. Using an in-domain corpus significantly improves the performance for a given task, whereas using a corpus in an unsuitable domain may decrease performance. These findings both indicate that an in-domain corpus improves the performance of word embeddings for the specific domains.

Problems in validation
To assess the quality (or usability) of the embeddings, validation methods are used. This are tasks designed to produce a metric that gives an indication of the usability of the provided embeddings. **?**] found that the validation method indicates only the quality of an embedding for a specific task. There is (yet) no methtod that can asses the usability of an embedding on all possible tasks, since each task may require other information to be embedded into the embedding. Validation methods use either labelled or unlabelled data. Labelled data is data that is in some way marked, so that the correct answer can be derived from it.

	WordSim	Men	RareWords
Best results from the research by Truong:	0.49	0.61	0.32
Average results from the research by Truong	0.45	0.59	0.32

Table 2: Results for the different validation sets of word simularity validations on domain specific texts from the study by **?**]

Unlabelled is the opposite, this data is not marked.
The usage of labelled data is common practice for validation methods, since the results produced by this data can be easily checked checked. Unpublished results of the study by Truong encounter this problem, they show high error rates on the validation scores, presented in Table 1. However, the word embeddings created correct document clusterings[**?**], this seems to indicate that the word-vectors are able to represent the words correctly but that the available validation sets cannot confirm this.
Furthermore, a study by **?**] found that the quality of embeddings are tasks specific, *different tasks favour different embeddings*. They also found that the embeddings encode information about word frequency, even in models that are created to prevent this. *This casts doubt on the common practice of using vanilla cosine similarity as a similarity measure.*
Therefore, we propose the validation of domain specific word-embeddings through a classification tasks, using multiple vector-distance calculations. This eliminates the need for labelled data in the validation of these domain specific word embeddings, will validate the quality of word embeddings for domain specific texts, and will validate the impact of different vector-distance measures on a categorization task.

Research Questions

- Have word-embeddings a higher accuracy for academic texts than TF-IDF for article classification?
- Which metric(s) can be used to measure the accuracy of word embeddings for scientific articles?