Short texts semantic similarity based on word embeddings

Karlo Babić, Sanda Martinčić-Ipšić, Ana Meštrović
University of Rijeka
Department of informatics
Radmile Matejčić 2, 51000 Rijeka
{kbabic, smarti, amestrovic}@domain.com

Francesco Guerra
University of Modena and Reggio Emilia
Department of Engineering Enzo Ferrari
via Vivarelli 10, 41125 Modena
francesco.guerra@unimore.it

Abstract. *Evaluating semantic similarity of texts is a task that assumes paramount importance in real-world applications. In this paper, we describe some experiments we carried out to evaluate the performance of different forms of word embeddings and their aggregations in the task of measuring the similarity of short texts. In particular, we explore the results obtained with two publicly available pre-trained word embeddings (one based on word2vec trained on a specific dataset and the second extending it with embeddings of word senses). We test five approaches for aggregating words into text. Two approaches are based on centroids and summarize a text as a word embedding. The other approaches are some variations of the Okapi BM25 function and provide directly a measure of the similarity of two texts.*

Keywords. semantic similarity, short texts similarity, word embeddings, *word2vec*, NLP

1 Introduction

Measuring semantic similarity of texts has an important role in the various tasks from the field of natural language processing (NLP) such as information retrieval, document classification, word sense disambiguation, plagiarism detection, machine translation, text summarization, etc. A more specific task, measuring semantic similarity of short texts is of great importance in applications such as opinion mining and news recommendation in the domain of social media (De Boom et al., 2016).

A large number of approaches have been developed for addressing this problem. Some of these approaches typically model short text as an aggregate of words and apply specific metrics to compute the similarity of these aggregations. Most of the existing techniques represent text as a weighted set of words (e.g., bag of words), where the order of the word in the text (i.e. the context) and the possible meanings associated to the words is not taken into account. Recently, neural networks have been adopted for building *word embeddings*, thus providing a real breakthrough to this field. Word embeddings represent a corpus-based distributional semantic model which describes the context in which a word is expected to appear. There is a variety of representation models based on word embeddings (Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017; Peters et al., 2018).

Within these models, the context of the words is taken into account in the process for defining the embeddings and the accuracy of the applications using them is typically improved. However, word embeddings have certain limitations: for example, they cannot capture more than one meaning per word (polysemy). Furthermore, a large number of lexical knowledge bases have been developed in the last years. The knowledge they convey could be exploited for creating embeddings that better represent the meaning of the words they are representing. For this reason, other techniques have been proposed to extend the aforementioned approaches with embeddings of words and meanings associated with words. This is the case of the *NASARI* dataset (Camacho-Collados et al., 2016) that integrates pre-trained word embeddings based on *word2vec* model with word senses embeddings reached from the BabelNet. BabelNet is a multilingual dictionary which contains synsets (Navigli and Ponzetto, 2012). It merges WordNet with other lexical and encyclopedic resources such as Wikipedia and Wiktionary.

Representing words in short texts with (semantic) embeddings is only the first step for capturing the meaning of them and being able to measure their similarities. Identifying the semantics of short texts is another challenging task due to the complexities of semantic interactions among words. More precisely, word embeddings can model the semantics of one word, but how to scale from words to texts is not a straightforward process. A large number of techniques have been proposed and there is no consensus in the community on how to proceed. The simplest approaches suggest taking the sum or the average (centroid) of the individual word embeddings for all words in the text. These approaches have been widely adopted in many experiments, for example, (Brokos et al., 2016; Rossiello et al., 2017; Sinoara et al., 2019) and in general, they perform well. However, by calculating only sum or centroid of a set of word embed-

This work has been supported in part by the University of Rijeka under the project: uniri-drustv-18-38

dings, we are losing a certain part of semantic information and thus maybe this is not an optimal approach. There are other possible approaches to generate text embeddings based on word embeddings like for example in (Kenter and De Rijke, 2015; Kusner et al., 2015).

SemTexT is project that involves University of Rijeka and University of Modena and Reggio Emilia with the aim of studying and developing semantic techniques for measuring the similarity of short texts. As one of the first actions in the project, the idea is to evaluate the performance of some of the existing techniques for representing short texts and measuring their similarities. In this paper, we describe our preliminary experiments where we evaluate how five similarity measures perform, with respect to human judgment. Two word representation models have been evaluated: one is a typical *word2vec* model; the second representation model is built on *NASARI* set, which includes word sense descriptions.

In short, in this paper, we address three main issues related to the task of measuring semantic similarity: (a) how to represent the words, (b) how to aggregate word representations for modeling short texts and (c) how to measure the similarity between aggregations. To resolve (a) we apply two existing representation models based on word embeddings; for (b) and (c) we test five methods that aggregate word embeddings and provide the semantic similarity score.

The results of our preliminary experiments in two datasets were quite surprising for us: the semantics provided by *NASARI* do not improve the performance in the results and centroid-based measures generally perform better than other more complex measures.

The rest of the paper is organized as follows. In Section 2, we present related work. In Section 3, we describe the approach with word sense embeddings and we give an overview of various word embeddings based methods for calculating semantic similarity of short texts. In Section 4, we provide evaluation results. Finally, in the last section, we give a conclusion and possible directions for future work.

2 Related Work

So far, there are numerous approaches developed for the task of measuring semantic similarity of words and texts that can generally be classified into two groups: corpus-based and knowledge-based (Mihalcea et al., 2006).

Knowledge-based measures of semantic similarity rely on external sources of knowledge (e.g. ontologies processed as semantic graphs or semantic networks, and/or lexical bases such as WordNet (Miller, 1995; Lenat, 1995), Wikipedia (Witten and Milne, 2008; Nastase and Strube, 2008), etc.). Moreover, these approaches use formal expressions of knowledge explicitly defining how to compare entities in terms of semantic similarity.

Corpus-based measures enable comparison of language units such as words, or texts based on statistics. They determine semantic similarity between words or texts using information derived from large corpora. These include traditional approaches like simple n-gram measures (Salton, 1989; Damashek, 1995), bag of words (Bow) (Salton et al., 1975; Manning et al., 2010) or more complex approaches such as Latent Semantic Analysis (LSA) proposed by Landauer (Landauer et al., 1998).

Recent trends in NLP prefer corpus-based approaches and representation models such as *word2vec* (Mikolov et al., 2013b), *GloVe* (Pennington et al., 2014), *FastText* (Bojanowski et al., 2017) and more recently *ELMo* (Peters et al., 2018).

The results of these models are words represented as embeddings with the property that semantically or syntactically similar words tend to be close in the semantic space (Collobert and Weston, 2008; Mikolov et al., 2013a).

Identifying the degree of semantic similarity of short texts based on the word embeddings is a challenging task that has been studied extensively in the past years. Certain approaches offer a sentence or document embeddings as a solution (Le and Mikolov, 2014; Cer et al., 2018). However, in this study, we are focused on the methods that enable determining semantic similarity of short texts based only on the word embeddings.

Mihalcea et al. proposed an approach for measuring semantic similarity of texts by exploiting the information that can be drawn from the similarity of the component words. The proposed approach is based on two corpus-based and six knowledge-based measures of word semantic similarity. According to the presented results, it outperforms the vector-based similarity approach in the task of paraphrase detection. However, their approach is rather traditional and it is not based on word embeddings.

Kusner et al. introduced a new measure, called the Word Mover's Distance (WMD), which measures the dissimilarity between two text documents (Kusner et al., 2015). Documents are represented using word embeddings and the distance is calculated as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document. The measure is evaluated in the task of text classification and the results show that it tends to have lower classification error rates in comparison to other state-of-the-art baseline methods. Furthermore, in (Brokos et al., 2016) authors used WMD method in the task of information retrieval. They apply the proposed method on the biomedical domain for retrieval documents from BIOASQ and proved that their method is competitive with PUBMED. These are examples of indirect evaluation of WMD method since the measure is not tested in the task of measuring semantic similarity. In our approach, we use similar methods and perform a direct

evaluation.

In (De Boom et al., 2016) authors defined a novel method for vector representations of short texts. The method uses word embeddings and learns how to weigh each embedding based on its idf value. The proposed method works with texts of a predefined length but can be extended. Authors showed that their method outperforms other baseline methods that aggregate word embeddings for modeling short texts.

Kenter and De Rijke proposed an approach for measuring short texts semantic similarity by combining word embeddings with methods based on external knowledge sources (Kenter and De Rijke, 2015). They used various text features to train a supervised learning algorithm. They employed a modification of Okapi BM25 function for document ranking in information retrieval and adjust it for the task of measuring semantic similarity of short texts. They showed that their method outperforms some baseline methods in task of measuring semantic similarity of short texts. In our approach, we test the same function for measuring semantic similarity. However, we perform evaluations using different word embeddings and different datasets. Moreover, we introduce two modified versions of the proposed method.

In the end, we briefly discuss approaches related to the Semantic Text Similarity, while a full description of extensive related work emerging from SemEval tasks in recent years is beyond the scope of this paper. In the 1st task of Semeval2015, the evaluation results in terms of Pearson correlation with the human judgments are around 0.5 and 0.6, with the exception of the highest result of 0.735, (Xu et al., 2015). In (Marelli et al., 2014) results of evaluation using Pearson correlation are in range from 0.479 to 0.828. As defined within the task of SemEval2014, these approaches used compositional distributional semantic models and other semantic systems on full sentences. However, these models are trained to fit this particular dataset and possibly not likely to hold in future tasks of the same kind.

3 Methodology

In this section, we describe the methodology adopted in the paper by introducing the embeddings and the methods used for the pairwise measurement of semantic similarity of short texts.

3.1 Word and Word Senses Embeddings

In the experimental evaluation, we use two sets of pre-trained word embeddings and compare the performance of two text representation models in the task of measuring semantic similarity.

The first model is based on classical *word2vec* embeddings. In particular, we test the $UMBC_{w2v}$, set of word embeddings trained on the UMBC corpus (Han et al., 2013). This is a set of word embeddings that are freely available and already used in a number of experiments in the literature. By using these embeddings we make our experiment more general and the results are comparable with other experimental evaluations. The application of this kind of embedding is straightforward: each word is replaced with its corresponding embedding from the *word2vec* set through a lookup table.

The second model is based on the *NASARI* set of embeddings. These embeddings incorporate external knowledge by introducing word senses embeddings through links to the BabelNet synsets (Camacho-Collados et al., 2016). In our experiments, similarly as in (Sinoara et al., 2019), we use a *NASARI* dataset combined with $UMBC_{w2v}$ embeddings, and we call this representation model as *NASARI+word2vec*. The application of these embeddings requires to use the Babelfy system (Moro et al., 2014) to retrieve the ID of the sense associated with each word. The ID is then used to find the embedding for that word or phrase in the *NASARI+word2vec* set of embeddings. For the word which is not covered in Babelfy and thus does not have any ID, the embedding is extracted from the $UMBC_{w2v}$ set. This way, we enable the disambiguation of different word senses.

Note that the sets of embeddings are both trained on the same vector space and their representation vector length is 300. Thus, all the embeddings are semantically comparable.

3.2 Methods for Measuring Short Texts Semantic Similarity

In this section, we introduce five methods used in our experiments for calculating semantic similarity scores between two short texts based on their words embeddings.

The first method is based on centroids. For a given text represented with the set of word embeddings V, the centroid of V is calculated according to the equation (1):

$$cent(V) = \frac{\sum_{v \in V} v}{|V|}. \quad (1)$$

The centroid is typically adopted in the literature for synthesizing the meaning of a text. We experiment also with a modified version of the centroid method that uses the inverse document frequency (idf) multiplied with each word vector. This variant builds weighted centroids, where the uncommon terms in the collection assume bigger importance:

$$cent_{idf}(V) = \frac{\sum_{v \in V} idf(v) \cdot v}{|V|}. \quad (2)$$

By using the centroids, the similarity measure of two documents sim_{cos} is computed as the cosine similarity between centroids of two texts t_1 and t_2 represented with sets of embeddings V_1 and V_2 respectively:

$$sim_{cos}(t_1, t_2) = cos(cent(V_1), cent(V_2)) \quad (3)$$

Analogously, sim_{cos2} is calculated as cosine similarity between weighted centroids of two texts (short texts or sentences).

We also experiment with three other methods that are based on the Okapi BM25 function. The first method has been introduced in (Kenter and De Rijke, 2015) and the other methods are simplified modifications of the original method.

The modified version of Okapi BM25 function that can be apply for measuring semantic similarity of two texts (short texts or sentences) introduced in (Kenter and De Rijke, 2015) is:

$$sts(t_l, t_s) = \sum_{w \in t_l} idf(w) \cdot \frac{sem(w, t_s) \cdot (k_1 + 1)}{sem(w, t_s) + k_1 \cdot (1 - b + b \cdot \frac{|t_s|}{avgtl})}, \quad (4)$$

where t_l is the longer text, and t_s is the shorter text. Variables $k1$ and b are parameters which can be optimised, variable $avgtl$ is average text length. Function sem for a given word w and text t is defined as:

$$sem(w, t) = max_{w' \in t} cos(w, w'). \quad (5)$$

Next, we introduce two modifications of equation (4) by leaving out constants $k1$ and b. The results are two simplified versions of equation 4. The rationale behind these simplifications are additional experiments that we perform by changing values of $k1$ and b (explained at the end of the fourth Section). Since there were no substantial differences in the results performed with different values of b and $k1$, we decided to remove those variables.

Equation (6) calculates average value returned by the function sem (5) multiplied by the idf.

$$sts_s(t_l, t_s) = \frac{\sum_{w \in t_l} idf(w) \cdot sem(w, t_s)}{|t_l|}. \quad (6)$$

Equation (7) is another modification of of (4). Here, instead of just calculating the idf of the word from the longer text, idf is calculated for words from both texts (w_s represents the word from the shorter text). One more difference is that the resulting value is passed through the log function so the most extreme values are reduced.

$$sts_{s2}(t_l, t_s) = log(\frac{\sum_{w \in t_l} (idf(w) + idf(w_s)) \cdot sem(w, t_s)}{|t_l|}). \quad (7)$$

4 Evaluation and Discussion

4.1 Datasets

To evaluate the performance of short texts similarity measures we used two datasets.

The first dataset (d1), called *SICK* dataset in its original version, is defined within the tasks of SemEval-2014 International workshop for the two tasks: determining the degree of relatedness between two sentences and detecting the entailment relation between sentences (Marelli et al., 2014). The dataset consists of 5,000 English sentence pairs. Each sentence pair is annotated with a score that represents the degree of sentence similarity according to a scale ranging from 1 to 5 (where 1 means that there is no semantic similarity and 5 refers to semantically equivalent sentences).

The second dataset (d2), refereed as a Lee dataset is defined in (Lee et al., 2005) for the task of evaluating measures for text to text similarity. The dataset is composed of 50 short English documents (sentences) presenting news from the Australian Broadcasting Corporations news mail service. Each document pair (2500 pairs in total) is annotated with the score of relatedness using discrete values from 1 to 5, proposed as an average score based on the proposal ten participants with an inter-agreement score of 0.61.

4.2 Evaluation Results

To evaluate and compare representation models and methods for measuring short texts semantic similarity (described in subsection 3.2), we computed the pairwise similarity of all short texts in both datasets. We compare the results with human judgments through the Pearson and Spearman correlations. Tables 1 and 2 show the results obtained against the *SICK* and the Lee datasets, respectively.

The rows represent the similarity measures experimented, namely sim_{cos}, sim_{cos2}, sts, sts_s and sts_{s2}. The columns represent the correlation measures computed with the *word2vec* embedding (a1) and *NASARI+word2vec* (a2).

The set of experiments performed on the *SICK* datasets show that according to the Pearson correlation, sim_{cos2} has the best performance in combination with *word2vec* model, while sts_{s2} has the best performance in combination with *NASARI+word2vec* model. Next, in the case of Spearman correlation, sim_{cos2} has the best performance with both representation models.

We repeat the same set of experiments on the Lee dataset and sim_{cos2} shows the highest values of correlations in all cases.

Overall comparison of two approaches: the first with *NASARI* embeddings and the second with *word2vec* embeddings shows that approach with classical *word2vec* embeddings slightly outperforms approach with *NASARI* embeddings on both datasets and

Table 1: Pearson (r) and Spearman (ρ) correlations of five similarity measures for word2vec (a1) and *NASARI+word2vec* (a2) approaches for the *SICK* dataset.

	r (a1)	ρ (a1)	r (a2)	ρ (a2)
sim_{cos}	0.642	**0.585**	0.586	**0.557**
sim_{cos2}	**0.661**	0.579	0.604	0.550
sts	0.503	0.468	0.470	0.443
sts_s	0.565	0.534	0.549	0.516
sts_{s2}	0.642	0.537	**0.612**	0.520

for all measures. This might be explained with the assumption that the Babelfy system does not achieve its full potential since it does not always return the correct word sense embedding for a given word within a context. There is certainly room for improvements of the BabelNet and Babelfy systems.

Table 2: Pearson (r) and Spearman (ρ) correlations of five similarity measures for *word2vec* (a1) and *NASARI+word2vec* (a2) approaches for the *Lee* dataset.

	r (a1)	ρ (a1)	r (a2)	ρ (a2)
sim_{cos}	0.582	**0.519**	0.472	0.464
sim_{cos2}	**0.589**	**0.519**	**0.502**	**0.478**
sts	0.283	0.193	0.276	0.225
sts_s	0.474	0.293	0.500	0.406
sts_{s2}	0.424	0.288	0.444	0.378

There are minor deviations compared to the results for the centroid similarity measure on the Lee dataset reported in Sinoara et al. (2019) because there is a new version of the *NASARI* dataset. However, the overall results of centroid based similarity are inline with the previous study for both approaches (*NASARI* approach and *word2vec* approach).

According to the Pearson correlation, we slightly outperform centroid measure with weighted centroid measure (on both datasets) and in some cases with variations of sts measure.

In comparison to the results reported for the *SICK* dataset, presented approaches are better than few approaches described in (Marelli et al., 2014). However, only *word2vec* in combination with both centroid-based methods slightly outperforms baseline (reported as an overlap of 0.63). For Lee dataset, the results show that both centroid-based methods and sts_{s2} perform better than human inter-agreement of 0.61.

Figure 1 shows scatter plots of relationships between automatically determined scores of semantic similarity and human ratings in four different cases. The first two cases (A) and (B) refer to the methods with the highest score on the *SICK* and Lee datasets respectively. Two other cases (C) and (D) illustrate relationships between automatic and human judgments in the worst cases. It is obvious from the large dispersion that automatically determined scores are not much correlated with the human intuition.

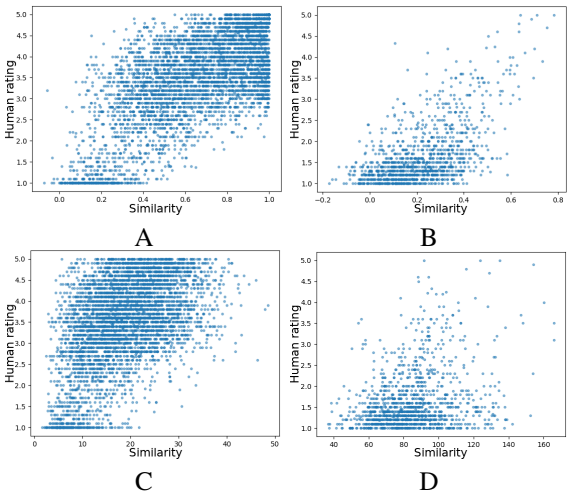

Figure 1: Scatter plots of relationships between human judgments and various approaches on two datasets: (A) the best results with sim_{cos} method on the *SICK* dataset; (B) the best results with sim_{cos2} method on the Lee dataset; (C) the worst results with sts method on the *SICK* dataset; (D) the worst results with sts method on the Lee dataset

Since sts measure has the lowest values for both Pearson and Spearman correlations for all cases, we did some extra experiments with tuning parameters $k1$ and b. Their default values are $k1 = 1.2$ and $b = 0.75$. Through tuning of those parameters, we found values that are optimal for this task. Parameter tuning and optimal values are shown in Figure 2.

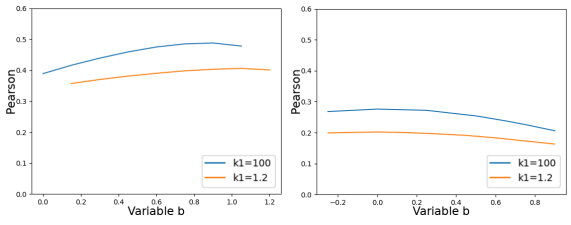

Figure 2: $k1$ and b tuning for datasets *SICK* and *Lee*. Y-axis is Pearson value and X-axis is b value. Top curve (blue) is for $k1 = 100$, and bottom curve (red) is for $k1 = 1.2$.

5 Conclusion

In this paper, we present preliminary research focused on the measuring semantic similarity of short texts. We test and compare two representation models: traditional *word2vec* model and its extension with em-

beddings of word senses *NASARI* provided by the Babelfy system. We combine these representation models with centroid-based and BM25-based methods and their variations.

Evaluation results on two datasets (*SICK* and Lee) in terms of Pearson and Spearman correlations indicate that *word2vec* model performs better than its extension. The reason might be that the *NASARI* dataset is not yet fully developed. We expect that with the better version of *NASARI* and Bablefy, this extended representation model, *NASARI + word2vec* will perform better.

Concerning the different methods for measuring similarity, there is no consensus on which method is the best. According to the Pearson correlation, weighted centroid measure slightly outperforms centroid measure (on both datasets) and variations of sts measure that we propose in this paper performs better than in sts general. Results are slightly above the human inter-agreement. The overall results are still not enough correlated with the human scores. However, there is still room for improvements.

For future work, we plan to systematically experiment with all other available representation models. Moreover, we will explore and incorporate more external knowledge resources like for e.g. ontologies, Google Knowledge Graph, Wikipedia, etc..

Acknowledgments

This work has been supported in part by the University of Rijeka under the project: uniri-drustv-18-38

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brokos, G.-I., Malakasiotis, P., and Androutsopoulos, I. (2016). Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. *arXiv preprint arXiv:1608.03905*.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848.

De Boom, C., Van Canneyt, S., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., and Weese, J. (2013). Umbc_ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52.

Kenter, T. and De Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420. ACM.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Lee, M. D., Pincombe, B., and Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Manning, C., Raghavan, P., and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Nastase, V. and Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. In *AAAI*, volume 8, pages 1219–1224.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rossiello, G., Basile, P., and Semeraro, G. (2017). Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 169.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., and Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971.

Witten, I. H. and Milne, D. N. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links.

Xu, W., Callison-Burch, C., and Dolan, B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.