Detecting Paraphrases for Portuguese using Word and Sentence Embeddings

Marlo Souza and Leandro M. P. Sanches

Institute of Mathematics and Statistics
Federal University of Bahia - UFBA, Salvador/BA, Brasil
msouza1@ufba.br,leandrompsanches@gmail.com

Abstract. Paraphrase identification is the task of determining whether two or more sentences of arbitrary length possess the same meaning. Methods to solve this task have many potential applications in Natural Language Processing systems. This work investigates the combination of different methods of sentence representation in a vector space model of language and linear classifiers to the problem of paraphrase identification for the Portuguese language.

Keywords: Paraphrase identification · Semantic textual similarity · Sentence embeddings.

1 Introduction

Paraphrase identification is the task of determining whether two or more sentences of arbitrary length possess the same meaning. Methods to solve this task have many potential applications in areas such as automatic summarization [21], information retrieval, question answering [26], automatic ontology construction [33], among others.

Recently, much work has been done in the area of paraphrase identification and the related task of semantic textual similarity [17, 32, 34]. Among the methods proposed in the literature, we can distinguish approaches based on lexical similarity measures, on contextual similarity measures and on distributional semantics.

Among those following the latter approach,much work has focused on what is commonly called *sentence embedding models*. A sentence embedding is a model that transforms a sentence of a given language into a vector in a high-dimensional vector space. Similar to word embeddings, it is supposed that the geometry of the vector space used to represent the sentences encodes important aspects of their meaning.

Sentence embeddings have been applied to many problems in Natural Language Processing, such as Machine Translation [5], Sentiment Analysis [22], Automatic Dialog Generation [34], etc. Particularly, for the English language, benchmarks for the task of measuring semantic similarity between sentences have become popular resources for evaluating the quality of sentence and word embedding models, e.g. the SICK dataset [25].

In this work, we perform some initial investigations on the application of sentence embedding methods for paraphrase identification for the Portuguese language.

2 Geometric representations of Words and Sentences

Word embeddings are models that explore the distributional similarity between words in a large *corpus* in order to learn representations of words of a language as points in a given high-dimensional vector space. Similarly, sentence embeddings are methods that aim to encode sentences as points in a given vector space in such a way as to preserve meaning.

Simple sentence representation models can be obtained by the composition, or *aggregation*, of the representations of the words composing the sentence in a given word embedding model. This method aims to explore the property of meaning compositionality, in which the meaning of a sentence is obtained by some transformation on the meaning of its constituents. Thus, methods following this approach [10, 27] aim to establish some transformation to perform such *aggregation*, i.e. they aim to learn how to compose the meaning of individual words to faithfully represent their contribution to the meaning of the sentence.

A trivial way to do so is to take the centroid of the words composing the sentence as its representation. This corresponds to the idea that each word contributes equally in determining the meaning of the sentence. This vector representation can be obtained taking the component-wise average of all the words in the sentence.

It is not clear, however, that each word contributes equally to the meaning of the sentence. In fact, some words may act as grammatical markers and their individual meaning may not contribute to the meaning of the sentence at all, e.g. the case of the word *pas* which has been grammaticalized in the verbal negation "*ne ... pas*" (not) in French. To account for the difference in the importance of each word to the meaning of the sentence, the sentence representation may be taken as the weighted aggregation (or pondered sum) of the vector of each word. Many different weighting strategies may be established in such a way as to take into consideration the structure of the sentence or the distributional properties of the words. A common approach, similar to Mihalcea *et al.*'s [27] approach to compute sentence similarity, is to take the Inverse Document Frequency (IDF) of each word on a given representative *corpus* as a measure of importance for the word. The idea is that the less common words contribute more - or have some *saliency* - in the meaning of the sentence.

Notice that most word embedding models aim to capture co-occurrence probabilities of words present in the training *corpus*. However, the presence of words out of context can cause noise in the trained model [3]. Thus, the method of simply aggregating word representations to compute the sentence representation may result in the accumulation of noise. To overcome this problem, Arora *et al.* [3] propose the use of matrix factorization methods to identify the principal component of the word vectors, which is interpreted as the *accumulated noise*.

This is then eliminated from the representation of the sentence. This technique is known as Smooth Inverse Frequency (SIF).

Works such as that of Kiros *et al.* [22], on the other hand, aim to learn the entire representation of a sentence from its distributional patterns in a large *corpus*, like the methods for word embeddings. These methods usually rely on the word representations and try to learn from the *corpus* the best way to aggregate such representations to compute the representation of the sentences.

Many different methods for learning sentence embeddings have been proposed in the literature, usually employing deep and recurrent neural networks in order to learn such representations [9, 22, 23, 29, 32]. These methods have been successfully applied to many *downstream applications* in NLP [8, 20, 24].

One of the most impactfull works on sentence embeddings is that of Kiros *et al.* [22], which proposes the skip-thought method. Skip-thought is an unsupervised method of learning sentence embeddings using an encoder-decoder architecture of neural networks [22] to predict the neighborhood of a certain sentence. The impact of such a method lies in the fact that it is unsupervised, and, thus, it does not require any annotated data, and can be re-used for many NLP applications. Supervised methods for sentence embedding such as InferSent [9], on the other hand, have proven to be successful for specific applications, but come with the price of relying on annotated data - which may not be available for all languages.

In this work, we will focus on the application of different sentence embedding models, and the related semantic similarity measures arising from these models, to the problem of paraphrase identification in Portuguese. In a sense, our work is similar to that of Feitosa and Pinheiro [14] or that of Fialho *et al.* [16], which evaluate the use of some lexical-based similarity measures to the problem of semantic textual similarity, restricted to the case of paraphrasing and using similarity measures arising from sentence embedding models.

3 Related Work

Work on paraphrase identification can be divided into three broad categories. First, there are the works based on heuristics such as semantic similarity measures and rich thesauri, such as [11, 15]. Other work, such as [31], compute contextual similarities, such as co-occurrence in a sentence or phrase, between words and explore such similarities to detect relatedness of meaning between two sentences - usually applying machine learning algorithms to identify the paraphrases. Finally, the third method relies on distributional semantics principles, such as the distributional hypothesis[1]

The work of Cordeiro *et al.* [11] proposes a metric for semantic relatedness between two sentences based on their overlapping of lexical units. Notice that works using lexical variation measures to identify paraphrases, such as [13], which use the Levenshtein distance between two sentences, are able to identify only

[1] The distributional hypothesis claims that linguistic items with similar statistical distributions in large *corpora* have similar meanings [30].

those examples in which the sentences have almost identical structure. While the work of Cordeiro *et al.* avoids many such pitfalls, as lexical overlap is a rather restrictive condition to identify paraphrases, their approach is limited in the sense that it cannot detect paraphrases in which there is significant variance in descriptions of entities and actions in the sentences, such as the use of different names and definite descriptions.

Works such as that of Mihalcea *et al.* [27] and that of Fernando and Stevenson [15], on the other hand, propose the exploitation of lexical similarity measures to identify paraphrases in the English language, based not only on lexical matching but on semantic, contextual or distributional similarities. These work explore rich information based on annotated thesauri, such as WordNet [28], and large *corpora*, as explored by Turney and Littman [35]. They are flexible in the sense that the can be employed using different similarity measures exploring either rich semantic resources or large unannotated *corpora* available for a language.

Socher *et al.* [32] employ recursive autoencoders (RAE), a kind of unsupervised deep neural network following the encoder-decoder model, to encode the tree structure of sentences. These representations are then applied to measure the word- and phrase-wise similarity between two sentences, which are dynamically pooled into a fixed length representation which is then used to train a paraphrase classifier.

Similarly, Yin and Schutze [36] propose the use of deep convolutional neural networks to solve the problem of paraphrase detection. They propose a new neural network architecture that, they claim, allows to encode multiple levels of granularity of the sentences meaning. These representations are then used to train a logistic classifier to identify paraphrases.

These more recent works are similar to that of Mihalcea *et al.* [27] and of Fernando and Stevenson [15] by also exploring distributional similarity in large unannotated *corpora* to compute semantic similarity between sentences. The difference is that the newer approaches consider the structure of the sentence to compute the semantic similarity between them, while those earlier ones do not take such information into consideration.

The work of Kiros et al [22] describes a model of unsupervised learning of a generic sentence encoder, which can be applied to different downstream tasks in NLP. Similar to what is done for word representation models, the authors train an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. The authors evaluate the generated models on 8 tasks: semantic relatedness, paraphrase detection, image-sentence ranking, question-type classification and 4 benchmark sentiment and subjectivity datasets.

Our work stems from these more recent works that apply deep neural networks and vector space models to represent the semantic information expressed in a sentence. We aim to evaluate their utility to the problem of paraphrase identification for the Portuguese language. Notice that other work on paraphrase identification for Portuguese has been conducted, especially on the context of the ASSIN joint evaluation for semantic similarity and textual inference [17].

While some of these works employ features obtained with word embeddings, most notably Hartmann [19], to our knowledge none of them evaluated different methods of sentence representation to this problem.

4 Using Representations of Sentences to Recognize Paraphrases

In this section, we describe the implementation paraphrase classifiers that receive two sentences and decide if they are examples of paraphrases. We investigate different linear classifiers trained on data of sentence representation and similarities obtained with using four different forms of sentence representation. Below we describe the data we use in our experiments as well as the results obtained in our investigation.

4.1 Data

In this work, we use three main data sources: a word embedding model for Portuguese, a non-annotated *corpus* of texts in the Portuguese language to train the Skip-Thought model and the ASSIN [17] *corpus* to train and evaluate our classifiers.

For the word embedding model used in our experiments, we chose to use the pre-trained FastText model for the Portuguese language from Facebook Research[2], which was trained on the *corpus* of Wikipedia articles written in Portuguese. We are aware that other word embedding models for Portuguese are available, particularly those in the NILC Word Embedding Repository analysed in the work of Hartman *et al.* [18]. Nevertheless, we chose the Facebook FastText model for two simple facts: first FastText has become one of the best performing models of Word Embeddings in the literature, see for example the experiments of Hartmann *et al.* [18] for the Portuguese language; seconde, the sizes of the best performing NILC models are simply too large, while Facebook model has a competing dimensionality, while still having a manageable size that allows us to perform our experiments.

The *corpus* used to train the Skip-Thought method is composed of 10,354,228 sentences and 308,261,905 tokens. The *corpus* was created taking all articles written in Portuguese from Wikipedia, an extract of around 1000 documents from the PLN-BR Full *corpus* of journalistic texts [7] and around 700 movie reviews from the websites *CinePlayers*[3] and *Cinema com Rapadura*[4].

To compute the weighted aggregated vector representations, as well as the SIF representations, we also used a dictionary of IDF values for words in the Portuguese language - both Brazilian Portuguese and European Portuguese variants - composed of 873,329 lexical units. This dictionary was obtained processing a fraction of *corpus* used to train the Skip-Thought model.

[2] Available at: `https://research.fb.com/fasttext/`

[3] `http://www.cineplayers.com`

[4] `http://cinemacomrapadura.com.br`

To train the classifiers, we used the train section of the ASSIN *corpus* [17] of textual similarity and paraphrases, limited to the Brazilian Portuguese variant of the *corpus*, composed of 2500 pairs of sentences annotated with sentence similarity and textual entailment relations, of which 116 are paraphrases. The classifiers were evaluated on the test section of the same *corpus*, containing 2000 pairs of sentences, from which 106 are positive examples of paraphrases.

4.2 Experiment

To evaluate the use of sentence embedding models on the problem of paraphrase detection in Portuguese, we trained a Skip-Thoughts model for the Portuguese language and applied this model, along with the Facebook FastText model, to compute different representations for each sentence pair. In this experiment, we employed the centroid representation (average of the word vectors), the weighted aggregation based on the IDF measure, the SIF representation, and the Skip-Thought representation of sentences.

We processed the data and obtained a different dataset for each sentence representation method containing the following features:

1. the vector representation $\overrightarrow{u}$ of the first sentence in the pair;
2. the vector representation $\overrightarrow{v}$ of the second sentence in the pair;
3. the component-wise product of vectors $\overrightarrow{u}$ and $\overrightarrow{v}$, i.e. the vector $\overrightarrow{u} \cdot \overrightarrow{v}$;
4. the norm of the vector $\overrightarrow{u} \cdot \overrightarrow{v}$;
5. the vector difference between vectors $\overrightarrow{u}$ and $\overrightarrow{v}$, i.e. $\overrightarrow{u} - \overrightarrow{v}$;
6. the norm of the vector $\overrightarrow{u} - \overrightarrow{v}$;
7. the cosine similarity between the two sentences;

Since the cosine similarity between two sentence vectors is regarded as able to encode some form of semantic similarity between them, we also created a different dataset consisting of the cosine similarity for each pair of sentences in the *corpus* using all different sentence representation methods investigated in this work. We wish to evaluate whether the sentence similarity can be used as an indicator for paraphrase. We also aggregated all the information into a single dataset, in which each point is composed of all information obtained for each representation method. We wish to evaluate with this dataset whether different representations can encode different aspects of the meaning of the sentences and whether these different aspects can be composed to identify paraphrases.

We evaluated the obtained classifiers using the well-established metrics of precision, recall and F1 for binary classification task [1].

4.3 Results

We trained different classifiers using data obtained with each sentence representation. In Table 1, we present the results obtained for each classifier explored in this work, i.e. Support Vector Machines (SVM), Naïve Bayes (NB), and Decision Tree (DT), trained on data obtained by each sentence representation method,

i.e. the average of the word vectors (Avg), the weighted aggregation of word vectors (Agg), the SIF representation (SIF) and the Skip-Thought representation (ST). Also, we trained the classifiers on a dataset containing only the obtained similarity values (Sim) and on all information combined (Total). Since the data are severely unbalanced, we also evaluated the performance of the classifiers with or without data balancing.

Table 1. Results of the evaluation of classifiers trained to identify paraphrase

Method	Classifier	Without Balancing			With Balancing		
		Prec	Rec	F1	Prec	Rec	F1
Avg	SVM	0.25	0.18	0.21	0.18	0.23	0.20
	NB	0.16	0.70	**0.26**	0.16	0.71	0.26
	DT	0.15	0.18	0.17	0.17	0.19	0.18
Agg	SVM	0	0	0	0	0	0
	NB	0.06	0.97	0.11	0.06	0.97	0.11
	DT	0	0	0	0.06	0.98	0.10
SIF	SVM	0	0	0	0.06	0.90	0.11
	NB	0.08	0.06	0.07	0.06	0.97	0.11
	DT	0	0	0	0.05	0.07	0.06
Skip	SVM	0.15	0.18	0.16	0.15	0.19	0.17
	NB	0.06	0.70	0.11	0.06	0.69	0.11
	DT	0.17	0.23	0.20	0.10	0.11	0.11
Sim	SVM	0	0	0	0.18	0.82	**0.29**
	NB	0.07	0.92	0.13	0.06	0.94	0.12
	DT	0.24	0.22	0.23	0.22	0.18	0.20
Total	SVM	0.24	0.18	0.21	0.24	0.20	0.22
	NB	0.07	0.70	0.13	0.07	0.69	0.13
	DT	0.14	0.15	0.15	0.20	0.21	0.21

In the unbalanced data, the classifier with the overall best performance was Naïve Bayes and the best representation was the average vector method. For the balanced data, both Naïve Bayes and Support Vector Machines classifiers have similar results, while Average vector representation and the similarity information achieved the best results.

5 Discussions

It is important to notice that the performance of the techniques investigated in this work is clearly below the performance reported for the English language (c.f. [22], for example) or those for textual inference reported by the competitors in the ASSIN challenge (c.f. [6] or [16]). One reason for this low performance may be due to lack of robustness of the word embedding model adopted, which has been trained on the *corpus* of Wikipedia articles - a small *corpus* for unsupervised learning of word embeddings. The evaluation of this model for semantic

similarity and analogies would help us better understand the obtained results. It is important to notice, however, that since the data are heavily unbalanced, the classifiers may suffer from overfitting on the positive examples - thus explaining the severe low precisions obtained.

Regarding the performance of both the weighted aggregated vector and the SIF representations, we notice that only around 393046 tokens in the vocabulary of the FastText model (composed of 592108 tokens) are in the IDF dictionary. This means that around 199062 tokens in the model have IDF value of 0, and thus have no effect in the sentence representation. This highlights that the different tokenization strategies adopted in our work and on the creation of the word embedding model may have impacted on the representations we achieved and, thus, on the results obtained.

It is also of notice that the performance of the skip-thought method may have suffered from the fact that the training *corpus* is relatively small compared to that used for the English language (composed of 74,004,228 sentences and 984,846,357 tokens).

It is interesting to remark that the top performing methods in our experiments were based on average vector representation and on semantic similarity measures between the encoded sentences. This means that the algebraic structure of the vector space may actually encode a great deal of information regarding the compositional semantics of sentences and that a simple model of sentence representation may be suitable to many downstream applications. These theoretical and empirical connections of word embeddings and compositional semantics, as well as the limitations of the encoder-decoder model, have been discussed before in the literature, notably by Arora and colleagues [2, 4, 12].

6 Final Remarks

This work investigated the application of different methods of sentence representation in a vector space model of language to the problem of paraphrase identification in the Portuguese language. While the results obtained for paraphrase classification were poor, compared to the results reported in the literature, we believe our results point to interesting avenues of investigation for paraphrase identification for the Portuguese language. Particularly, simple sentence representation methods and classifiers, namely average vector and semantic similarity representations and Nave Bayes classifier, obtained the best results, indicating that a great deal of semantic information of the sentences are encoded in the geometry of the word representation models.

As a future work, we intend to run experiments using the word embedding models created by Hartmann *et al.* [18] for the Portuguese language, which have a well-studied performance for the tasks of semantic similarity and of analogy identification. Also, we intend to train the skip-thoughts model on a larger *corpus* and re-evaluate its performance.

References

1. Alpaydin, E.: Introduction to Machine Learning. MIT Press (2009)
2. Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: Linear algebraic structure of word senses, with applications to polysemy. Transactions of the Association of Computational Linguistics **6**, 483–495 (2018)
3. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: Proceedings of the 7th ICLR. (2017)
4. Arora, S., Risteski, A., Zhang, Y.: Do GANs learn the distribution? some theory and empirics. In: Proceedings of the 8th ICLR (2018)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computing Research Repository **abs/1409.0473** (2014), `http://arxiv.org/abs/1409.0473`
6. Barbosa, L., Cavalin, P., Guimaraes, V., Kormaksson, M.: Blue man group no ASSIN: Usando representações distribuídas para similaridade semântica e inferência textual. Linguamática **8**(2), 15–22 (2016)
7. Bruckschen, M., Muniz, F., de Souza, J.G.C., Fuchs, J.T., Infante, K., Muniz, M., Gonçalves, P.N., Vieira, R., Aluísio, S.: Anotação lingüística em XML do corpus PLN-BR. Tech. rep., Universidade de São Paulo, São Paulo, Brazil (2008)
8. Cer, D.t.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
9. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the EMNLP 2017. pp. 670–680. Association for Computational Linguistics (2017)
10. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single vector: Probing sentence embeddings for linguistic properties. Computing Research Repository **abs/1805.01070** (2018), `http://arxiv.org/abs/1805.01070`
11. Cordeiro, J., Dias, G., Brazdil, P.: A metric for paraphrase detection. In: Proceedings of the 2nd ICCGI. IEEE (2007)
12. Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S.J., Goodman, N.D.: Evaluating compositionality in sentence embeddings. Computing Research Repository **abs/1802.04302** (2018), `http://arxiv.org/abs/1802.04302`
13. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th CICLing. p. 350. Association for Computational Linguistics (2004)
14. Feitosa, D., Pinheiro, V.: Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual. In: Proceedings of the 11th STIL. pp. 161–170 (2017)
15. Fernando, S., Stevenson, M.: A semantic similarity approach to paraphrase detection. In: Proceedings of the 11th CLUK. pp. 45–52 (2008)
16. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESC-ID@ ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. Linguamática **8**(2), 33–42 (2016)
17. Fonseca, E., Santos, L., Criscuolo, M., Aluísio, S.: ASSIN: Avaliacao de similaridade semantica e inferencia textual. In: Proceedings of the 12th PROPOR. pp. 13–15 (2016)
18. Hartmann, N., Fonseca, E.R., Shulby, C., Treviso, M.V., Rodrigues, J., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Computing Research Repository **abs/1708.06025** (2017), `http://arxiv.org/abs/1708.06025`

19. Hartmann, N.S.: Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. Linguamática **8**(2), 59–64 (2016)
20. Howard, J., Ruder, S.: Fine-tuned language models for text classification. Computing Research Repository **abs/1801.06146** (2018), http://arxiv.org/abs/1801.06146
21. Jing, H., McKeown, K.R.: Cut and paste based text summarization. In: Proceedings of the 1st NAACL. pp. 178–185. Association for Computational Linguistics (2000)
22. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-thought vectors. Computing Research Repository **abs/1506.06726** (2015), http://arxiv.org/abs/1506.06726
23. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st ICML. pp. 1188–1196 (2014)
24. Logeswaran, L., Lee, H.: An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893 (2018)
25. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th SemEval. pp. 1–8 (2014)
26. Marsi, E., Krahmer, E.: Explorations in sentence fusion. In: Proceedings of the 10th ENLG. pp. 109–117 (2005)
27. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: 21st AAAI. pp. 775–780. AAAI Press, Boston (2006)
28. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
29. Patro, B.N., Kurmi, V.K., Kumar, S., Namboodiri, V.P.: Learning semantic sentence embeddings using pair-wise discriminator. arXiv preprint arXiv:1806.00807 (2018)
30. Sahlgren, M.: The distributional hypothesis. Italian Journal of Disability Studies **20**, 33–53 (2008)
31. Shinyama, Y., Sekine, S., Sudo, K.: Automatic paraphrase acquisition from news articles. In: Proceedings of the 2nd HLT. pp. 313–318. Morgan Kaufmann Publishers Inc. (2002)
32. Socher, R., Huang, E.H., Pennington, J., Ng, A.Y., Manning, C.D.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Proceedings of the 24th NIPS. pp. 801–809. Curran Associates Inc., USA (2011)
33. Subhashree, S., Kumar, P.S.: Enriching linked datasets with new object properties. Computing Research Repository **abs/1606.07572** (2016), http://arxiv.org/abs/1606.07572
34. *et al*, Y.Y.: Learning semantic textual similarity from conversations. Computing Research Repository **abs/1804.07754** (2018), http://arxiv.org/abs/1804.07754
35. Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Computing Research Repository **cs.LG/0212012** (2002), http://arxiv.org/abs/cs.LG/0212012
36. Yin, W., Schütze, H.: Convolutional neural network for paraphrase identification. In: Proceedings of the NAACL HLT 2015. pp. 901–911 (2015)