Natural Language Processing
A study of Language Models using word2vec

Isabel B. Amaro, Adriano Veloso

[1]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil

{isabel.amaro,adrianov}@dcc.ufmg.br

Abstract. *Language Models are probability distributions over sentences. They are used in Natural Language Processing to retrieve information from non-structured corpus. With the advance of Internet and social networks, thousands of data are constantly generated, making this field a great corpus for NLP studies. This work analyzes two language models implementations, Continuous Bag-Of-Words and Skip-Gram, in a small Wikipedia corpus.*

1. Introduction

With the popularization of social networks, a massive amount of data is constantly being generated and growing exponentially in the last few years [Cambria and White 2014]. This makes the content-sharing services a very rich non structured corpus, with opinions, ideas, etc [Cambria and White 2014]. Natural Language Processing is a Computer Science field that studies the best ways to retrieve, process and generate from representations of human language [Cambria and White 2014]. Machine translation, information retrieval, text summarization, question answering, information extraction, topic modeling, and more recently, opinion mining, are some focus of NLP researches [Cambria and White 2014]. To do this, Language Models are used in NLP as probability distributions over sentences. A well known Language Model is word2vec, which represents each word from a corpus as a vector. Each word in the space is related to a context, which are the other words that are around it [Mikolov et al. 2013].

This work explains two implementations of word2vec: Continuous Bag-Of-Words and Skip-Gram. Then, it analyses these two implementations varying the corpus' size and context window size.

2. word2vec

Word2vec is an unsupervised learning method capable of representing words in a vector space, using a neural network to learn predict the neighbors within a given text window for each word in the vocabulary [Jansen 2017]. Two word2vec language models are Continuous Bag-Of-Words Model and Skip-Gram Model. The Continuous Bag-Of-Words uses continuous distribution representation of the context to predict the current word. Thus, as standard bag-of-words model, the other of words in context doesn't change the prediction [Mikolov et al. 2013]. On the other hand, in Skip-Gram, the context is the one to be predicted given the current word. The Skip-Gram model is capable get better results increasing the range of the context, but it is related to more computational cost [Mikolov et al. 2013].

3. Trained models

To do this experiment, a small Wikipedia corpus was used to train CBOW and Skip-Gram models, varying the context window sizes. The used corpus was also split into smaller sizes to evaluate its impact in final results. The word2vec code was also provided by Google.

4. Result and analysis

To evaluate the model language, 19544 sentences were offered as input to the algorithm. The sentences are a sequence of words, and the algorithm must predict correctly the next word. The cosine distance was used to calculate the distance between the predicted word and the expected word.

Table 1: Example of the model language evaluation

Sentence	Expected word	Predicted Word	Cosine distance
decrease decreasing fly	flying	flies	0.03

For both models, the parameters the window size and corpus size were set the same way, so we could see better each algorithm's performance. In CBOW, the best performance was related to smaller window size of context, producing smaller average distance between the predicted and expected word. On the other hand, Skip-Gram works better for medium window sizes. For this situation, the accuracy is not the best evaluation metric, as the corpus is too small and poor. This is proved by the increase of accuracy as increasing the vocabulary size for both models.

The best trained model obtained was Skip-Gram, with context sixe of eight words and trained with full corpus. The word embeddings for this model can be visualized in Figure 1. The word embeddings show the correlation of distances between words.

Table 2: CBOW results

Vocab. Size	Window	Best Distance	Worst Distance	Average Distance	Accuracy
39071	4	0.0	0.67	0.1	0.25
39071	8	0.0	0.77	0.1	0.27
39071	16	0.0	0.72	0.11	0.28
57014	4	0.0	0.68	0.07	0.41
57014	8	0.0	0.6	0.08	0.43
57014	16	0.0	0.65	0.09	0.42
71291	4	0.0	0.63	0.07	0.5
71291	8	0.0	0.65	0.07	0.54
71291	16	0.0	0.72	0.08	0.51

Table 3: Skip-Gram results

Vocab. Size	Window	Best Distance	Worst Distance	Average Distance	Accuracy
39071	4	0.0	0.5	0.1	0.2
39071	8	0.0	0.52	0.1	0.27
39071	16	0.0	0.63	0.11	0.31
57014	4	0.0	0.47	0.08	0.36
57014	8	0.0	0.49	0.08	0.43
57014	16	0.0	0.49	0.09	0.43
71291	4	0.0	0.51	0.07	0.48
71291	8	0.0	0.47	0.07	0.52
71291	16	0.0	0.51	0.08	0.5

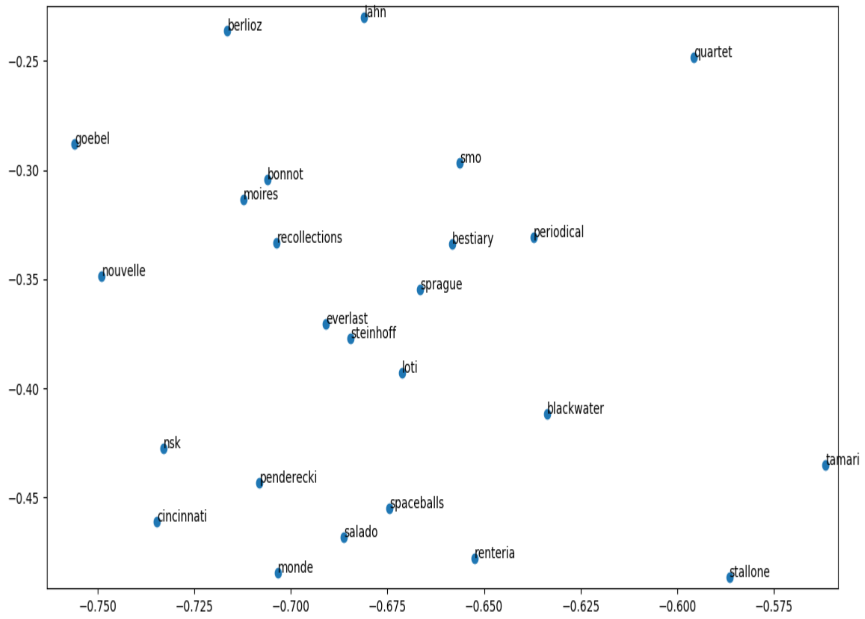

Figure 1: Zoom in word embeddings from Skip-Gram model with maximum vocabulary size and window size 8

5. Conclusion

This work is a case of study of Continuous Bag-Of-Words and Skip-Gram models. A small Wikipedia corpus was used to train the models with word2vec code provided by Google, varying the corpus and context sizes for both. The results show CBOW has better performance with smaller context sizes, while Skip-Gram has better performance with medium context sizes. The predominant evaluation metric is the average distance between the predicted and expected result, because the corpus is not big enough to generate a well trained model, which produces very low accuracy. Despite this, this work fulfill its goal which is analyze CBOW's and Skip-Gram's behaviour with same arguments.

References

Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Comp. Int. Mag.*, 9(2):48–57.

Jansen, S. (2017). Word and phrase translation with word2vec. *CoRR*, abs/1705.03127.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.