An Ensemble Method for Producing Word Representations focusing on the Greek Language

Michalis Lioudakis
Athens University of
Economics and Business
Greece
mlioudakis@hotmail.com

Stamatis Outsios
Athens University of
Economics and Business
Greece
soutsios@aueb.gr

Michalis Vazirgiannis
Athens University of
Economics and Business
Greece,
Ecole Polytechnique
France
mvazirg@aueb.gr

Abstract

In this paper we present a new ensemble method, Continuous Bag-of-Skip-grams (CBOS), that produces high-quality word representations putting emphasis on the modern Greek language. The CBOS method combines the pioneering approaches for learning word representations: Continuous Bag-of-Words (CBOW) and Continuous Skip-gram. These methods are compared through intrinsic and extrinsic evaluation tasks on three different sources of data: the English Wikipedia corpus, the modern Greek Wikipedia corpus, and the modern Greek Web Content corpus. By comparing these methods across different tasks and datasets, it is evident that the CBOS method achieves state-of-the-art performance.

1 Introduction

Neural networks have significantly affected Natural Language Processing (NLP) tasks. One of those tasks is representation learning for words, also known as word embeddings that represent words/tokens in a low dimensional Hilbert space where similarity computations are feasible and enable machine learning algorithms. The main idea behind word embeddings is the distributional hypothesis (Harris, 1954), which states that the meaning of a word can be captured by the context in which it appears.

Word embeddings are beneficial for most NLP applications increasing the overall performance and capturing different aspects of similarity among words. Numerous researches have shown these benefits in sequence tagging (Ma and Hovy, 2016; Lample et al., 2016) and text classification (Kim, 2014). Recently, Qi et al. (2018) have shown that pretrained word embeddings may be a valuable feature in machine translation, particularly in low-resource scenarios.

While living in the NLP era passing from static word representations to dynamic (contextualized) word representations, there are still applications where static word embeddings (word2vec, fastText, GloVe) are used, such as various RNNs/CNNs models. It is also known that in various NLP tasks, using a concatenation of context-aware word embeddings with static word embeddings (Peters et al., 2018a; Akbik et al., 2018) achieves better results.

We propose a new architecture **Continuous Bag-of-Skip-grams (CBOS)**, aiming to combine the benefits from Skip-gram and CBOW approaches. Our model achieves competitively high accuracy across different tasks compared to the aforementioned models. These results lead to an overall increased performance of the word embeddings. In addition, the CBOS architecture does not increase the computational cost significantly due to its efficient implementation. Thus CBOS can be trained on vast amounts of text corpora within a reasonable time.

The main contributions of our work are:

- Continuous Bag-of-Skip-grams (CBOS), a new ensemble word embeddings method
- Two new modern Greek language resources (a dataset for the classification task and a dataset for the NER task)
- A comprehensive comparative evaluation of CBOS, CBOW and Skip-gram models trained on three datasets, in two different languages.

The rest of the paper is organized as follows: Firstly, section 2 is a brief overview of previous work that has been done on word embeddings and NLP in modern Greek. Section 3 describes the data and tools that were used or produced for the training of our model. In section 4, our proposed

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 99–107
Decmber 04, 2020. ©2020 Association for Computational Linguistics

CBOS model is explained along with its differences to other popular models. Section 5 presents the evaluation methods used for comparing models in the experimental setup and Section 6 shows the results of the different experiments. Finally, in section 7 we provide conclusions based on the results of the experiments.

2 Previous Work

2.1 Static word embeddings

Two of the most popular approaches to produce static word vectors are the Skip-gram and the Continuous Bag-of-Words (CBOW) architectures, as implemented in word2vec (Mikolov et al., 2013a) and fastText (Bojanowski et al., 2017). The Skip-gram model predicts nearby words given a source word, while the CBOW model, predicts the source word according to its context. The latest version of these models is enriched with subword information in order to overcome some of their shortcomings (Bojanowski et al., 2017).

Even though these two methods produce high-quality word representations, each method achieves the highest accuracy in distinct categories of the word analogy questions. More precisely, the Skip-gram method performs better in semantic categories, while the CBOW method outperforms Skip-gram in syntactic tasks (Mikolov et al., 2013a). Our newly proposed method tries to benefit from both categories in order to increase the overall accuracy.

2.2 Contextualized word embeddings

Recent work in the area have shown that contextualized word embeddings outperform traditional word embeddings. This new class of embeddings proposes the production of various representations of each word based on its context, and not a single global representation. Embeddings from Language Models (ELMo) (Peters et al., 2018b) is one of these approaches and it is based on the representations obtained by a bidirectional language model. Devlin et al. (2018) introduced Bidirectional Encoder Representations from Transformers (BERT) which utilizes a deep language model based on a Transformer network.

2.3 Word embeddings evaluation

Concerning the comparison of the word representation models, many studies have been focused on word embedding evaluation (Ghannay et al., 2016; Wang et al., 2019; Schnabel et al., 2015). These studies have examined the intrinsic quality of word embeddings, along with their impact when used as inputs in other NLP application tasks. Thus, we evaluate the word representation methods in two different kind of evaluation tasks: intrinsic and extrinsic evaluation.

2.4 Resources in modern Greek language

It is widely known that modern Greek (hereafter simply Greek) resources are limited, especially compared to other rich-resource languages (e.g. English, French, German). Despite the work of Outsios et al. (2018, 2019) that published a large dataset crawled from millions of Greek webpages and an evaluation framework for Greek word embeddings, Greek language continues to be considered as a low-resource language. One of the most recent work in Greek NLP has been published by Koutsikakis et al. (2020), where the authors introduced GREEK-BERT, a monolingual BERT-based language model for Greek. Aim of the present work is to enrich the publicly available resources for the Greek language.

3 Data Sources and Tools

In this section, we describe the datasets that were used/produced for this research, along with their sources. Furthermore, we present the tools and libraries that were used for the development of word embeddings models.

3.1 Wikipedia Corpus

Wikipedia is the largest, with content in more than 200 distinct languages, free online encyclopedia. The quality standards followed by authors and the rigorous revisions by editors of the Wikipedia community are ensuring that the articles are of high quality. It has been used in various tasks, among others in information extraction (Wu and Weld, 2010) or word sense disambiguation (Mihalcea, 2007).

In this paper, we used the first 10^9 bytes of the English Wikipedia dump on March 3, 2006 provided by Matt Mahoney[1]. The data is UTF-8 encoded XML consisting primarily of English text. The English Wikipedia corpus contains 243K article titles. The primary preprocessing step was to extract the text content from the XML dumps. For this purpose, the script wikifil.pl was used as published by Matt Mahoney. The final preprocessed file consists from 680MB of text data and 124M words.

In addition, the Greek Wikipedia dump from December 2018 was used for training. A few basic preprocessing steps were implemented. These steps included lowercasing of all words and removing punctuation. The finalized text file used for training contains 800MB of text data and 68M words.

3.2 Greek Web Content Corpus

Recently, Outsios et al. (2018) have collected and crawled the most extensive Greek corpus available from about 20M URLs with Greek language content. First, the Greek corpus was extracted in Web Archive (WARC) format and then several pre-processing and extraction steps were applied. This process has produced a single uncompressed text which was used by our work. Greek language n-grams were also offered. Some details for the Greek corpus are listed below:

- Raw crawled text size: 10TB
- Text after pre-processing size: 50GB
- |Tokens|: 3B
- |Unique sentences|: 120M
- |Unigrams|: 7M
- |Bigrams|: 90M
- |Trigrams|: 300M

3.3 New Greek Datasets

One of our contributions in this work is the production of two new datasets for the text classification and NER tasks for the Greek language. These datasets will be publicly available[2].

3.3.1 Text Classification dataset

For the Greek classification task we produced a new dataset from newspaper Makedonia[3]. The full dataset contains 8005 articles from categories like Sports, Reportage, Economy, Politics, International, Television, Arts-Culture, Letters, Opinions etc. For the experiments the top seven categories are selected as a balanced dataset.

3.3.2 NER dataset

For the Greek NER task we produced a new dataset in CoNLL-2003 format, from Spacy's Greek ner.jsonl [4].

3.4 FastText Library

FastText is an open-source library that allows users to learn text representations and text classifiers. It supports training Continuous Bag-of-Words or Skip-gram models using different loss functions and a variety of tuning parameters.

Our contribution to fastText Library is the CBOS method that can be used for training. The source code will be made publicly available[5].

4 Proposed Model

4.1 Continuous Bag-of-Skip-grams

The new model, Continuous Bag-of-Skip-grams (CBOS), proposed by this work, is a combination of CBOW and Skip-gram models and was named respectively. The main idea behind CBOS is that, given a word *w* and a context window *c*, the training should capitalize on both training techniques in order to combine their benefits. So we consider two training phases:

i. A phase where *w* is trained by predicting every word in the context window *c* (Skip-gram).

ii. A phase where a bag-of-words is created from all words in the context window *c*, except a randomly selected word *p* which is used for predicting and word *w* which was used for training in the previous phase (CBOW).

Thus, if *D* is the set of correct word-context pairs, the probability functions of the two phases can be defined as follows:

$$P(D=1|w,c_{1:k}) = \prod_{i=1}^{k} \frac{1}{1+e^{-w\cdot c_i}} \tag{1}$$

$$P(D=1|p,c_{1:k}(p,c_i \neq w)) = \frac{1}{1+e^{-(p\cdot c_1+\ldots+p\cdot c_k)}} \tag{2}$$

[1] http://mattmahoney.net/dc/textdata.html

[2] http://archive.aueb.gr:7000/resources/

[3] http://www.greek-language.gr/greekLang/modern_greek/tools/corpora/makedonia/content.html

[4] https://github.com/eellak/gsoc2018-spacy/blob/dev/spacy/lang/el/training/datasets/annotated_data/ner.jsonl

[5] https://github.com/mikeliou/greek_word_embeddings

It is essential to note here, that we selected our proposed CBOS architecture between 6 different implementations, based on accuracy performance in the Greek word analogy task (Section 6.1). Furthermore, the CBOS method includes every feature and tuning parameter proposed by (Mikolov et al., 2013a,b; Bojanowski et al., 2017) as implemented in fastText Library (e.g. subword information, negative sampling).

As a working example, consider the sentence *"I am reading a paper about word embeddings"* with a window of 2 words before and after the current word. The current word for the first phase of training is "paper" and the randomly selected word to predict in the second phase is "about". In the first phase, "paper" will make four predictions, one for each word in the context window ("reading", "a", "about", "word"). In the next phase, every word vector, except the one selected randomly ("about") and the one used for training in the previous phase ("paper"), will be summed in a unique vector and will predict the word "about". This example is illustrated in Figure 1.

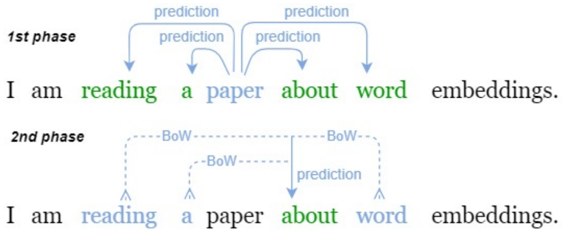
1st phase
prediction
prediction
prediction
prediction
I am reading a paper about word embeddings.
2nd phase
BoW
BoW
BoW
prediction
I am reading a paper about word embeddings.

Figure 1: A visualization of the CBOS model.

This simple step added to the training of each word seems vital for the improvement of the quality of word embeddings. The additional complexity by this step does not change the complexity class of the algorithm as it appears below (Table 2) in the execution times of the different models.

As shown, the CBOS model has a few more iterations on the training data than Skip-gram and CBOW models due to the second phase of training. In order to have a fair comparison between the different models in Section 6, CBOW and Skip-gram models were also trained with double the epoch size (10 instead of 5).

5 Evaluation

The quality of our new embedding method was thoroughly evaluated on the basis of restricted lexical semantics tasks, such as scoring word similarity and linear relationships for analogies (intrinsic evaluation). In order to focus on how performance correlates with downstream NLP tasks, two characteristic tasks are selected: a sequence labelling task for word level and a text classification task for sentence level (extrinsic evaluation).

5.1 Intrinsic Evaluation

Intrinsic tasks evaluate the quality of word representations generated by an embedding technique. These tasks measure syntactic or semantic relationships between words and, typically, are fast to compute. In this work, three different intrinsic evaluation tasks are selected: word analogy, outlier detection and word similarity.

5.1.1 Word Analogy

A common way of evaluating word embeddings is using the vectors produced to predict syntactic and semantic connections like "king is to queen as father is to ?".

Mikolov et al. (2013b) were the first to utilize word analogy as a method of creating connections between words, by using the offset of their vectors. The evaluation of word analogy is based upon the observation that simple arithmetic operations in a word vector space can reveal semantic and syntactic relationships between words: given the three words, a, b and c, the task is the identification of the word d, so that the relationship $c{:}d$ is the same as the relationship $a{:}b$ (Pereira et al., 2016; Turian et al., 2010). The evaluation dataset published by Mikolov and colleagues was used for the evaluation of the English word embeddings in this work.

An evaluation framework for the Greek word embeddings has recently been introduced by (Outsios et al., 2019). This evaluation framework focuses on intrinsic evaluation which evaluates the trained word embeddings using semantic and syntactic analogies and especially in word similarity and word analogy. In this work, for the evaluation of Greek word embeddings, we use the word analogy dataset[6].

5.1.2 Outlier Detection

Outlier detection task is a relatively new task for evaluating word representations and was proposed by (Camacho-Collados and Navigli, 2016). The goal is to distinguish the unrelated word in a group

[6] `http://archive.aueb.gr:8085/files/questions_greek.txt`

of words. This task evaluates the ability of vector space models to form semantic clusters in order to distinguish the outlier word. Furthermore, Camacho-Collados and Navigli (2016) define Outlier Position Percentage (OPP) which considers the position of the outlier in the group of words ranked by the compactness score.

5.1.3 Word Similarity

Word similarity is used for evaluating the distance between word vectors and semantic similarity perceived by humans. The goal of the word similarity task is to evaluate how accurately the human perceived similarity was captured by the word representations. The most common metric used in this evaluation is cosine similarity.

5.2 Extrinsic Evaluation

Extrinsic tasks are used to evaluate the contribution of word representations in the performance of a model in any downstream NLP task. In this work, we chose two different extrinsic evaluation tasks: text classification and named entity recognition.

5.2.1 Text Classification

Text classification is one of the most widely used NLP tasks. The goal of this task is to predict the class of the given text. The datasets used for the evaluation of text classification task were: the AG News dataset for the English language[7]and the contributed dataset referenced in Section 3.3.1 for the Greek language.

5.2.2 Named Entity Recognition

The named entity recognition (NER) task focuses on locating and classifying information units into pre-defined categories. For example, such categories can be person names, organizations, locations and time expressions. The datasets used for the evaluation of NER task were: the CoNLL-2003[8]for the English NER dataset and the contributed dataset referenced in Section 3.3.2 for the Greek language.

6 Experimental Results

6.1 Alternative CBOS implementations

Before we culminate in the CBOS model proposed earlier, we implemented different versions of CBOS in order to achieve the highest accuracy to the Greek word analogy task. The different versions of CBOS are described below:

- *Next-word incremental CBOS*: After the first phase of predictions, the bag-of-words is formed incrementally starting from the first word at the left. After the addition of each word to the bag-of-words, a prediction is made on the next word.

- *Central-word incremental CBOS*: The same process as in previous method is followed but, instead of predicting the next word, the prediction is made on the central word of the window.

- *Non-random CBOS*: This implementation follows the same steps of CBOS except for the randomly chosen word in the second phase. The chosen word for prediction is the central word.

- *Variable context window CBOS*: In the second phase of CBOS, the context window is changed to a random number between 1 and 5. Thus, the bag-of-words could contain different words used for the second phase of training.

- *Non-repeated words CBOS*: This method does not add any word to the bag-of-words that is already contained.

Model	**Semantic**	**Syntactic**	**Total**
Baseline	**52.72**	48.23	**50.16**
Next-word incremental	43.35	**52.62**	48.63
Central-word incremental	9.27	36.67	24.90
Non-random	37.96	50.60	45.17
Variable context window	48.49	47.49	47.92
Non-repeated words	51.12	47.65	49.14

Table 1: Accuracy of the different CBOS versions on word analogy task using the Greek Wikipedia Corpus for training.

For the comparison presented in Table 1, the Greek Wikipedia dataset and the default parameters were used for training. For the evaluation, the closest vector is evaluated and the out-of-vocabulary (OOV) words are excluded.

6.2 Training time

Before comparing the evaluation scores across the various tasks, in Table 2 we present the training

[7] https://www.kaggle.com/amananandrai/ag-news-classification-dataset

[8] https://www.kaggle.com/alaakhaled/conll003-englishversion

time of each model since the computational cost is a critical factor.

Model	English Wikipedia	Greek Wikipedia	Greek Web Content
CBOW ep10	20m 14.019s	16m 25.623s	791m 46.704s
CBOW ep5	**10m 14.099s**	**8m 24.453s**	**399m 7.573s**
Skip-gram ep10	29m 27.670s	22m 39.659s	1395m 20.622s
Skip-gram ep5	14m 11.977s	11m 45.747s	589m 39.222s
CBOS ep5	21m 10.789s	16m 55.784s	810m 43.196s

Table 2: Training time of the CBOS model and baselines across different datasets.

6.3 Intrinsic Evaluation

6.3.1 Word Analogy

For the first evaluation, the English Wikipedia dataset was used. The three models were trained using the default parameters provided by the fastText library and were evaluated using the word analogy task for English language (Mikolov et al., 2013a). Only the closest vector (top-1) is considered for a successful prediction. The out-of-vocabulary (OOV) words are excluded. Results are presented in Table 3.

The CBOS model does not achieve the highest accuracy in either the semantic or the syntactic category, but it outperforms the other two models trained with the same epochs in the total score. The CBOW model trained on 10 epochs achieves the best accuracy in the syntactic category. The Skip-gram ep10 model outperforms the other two in the semantic category and the total score but has the worst execution time.

The next two evaluations used the Greek Wikipedia dataset and the Greek Web Content dataset for the training of the three models. Every model was trained using the default tuning parameters suggested by the FastText framework. The word analogy task for the Greek language (Outsios et al., 2019) was used for the evaluation of the closest vector, and the OOV words were not evaluated. The results for the Greek Wikipedia dataset and the Greek Web Content corpus are shown in Table 3.

Concerning the Greek Wikipedia dataset, the CBOS approach achieves the highest accuracy in all categories compared to the models trained on 5 or 10 epochs. The Skip-gram method trained with 10 epochs achieves the highest accuracy in the semantic category, but the CBOS method outperforms the other two methods in the syntactic category and total accuracy.

The results related to the Greek Web Content dataset show that the CBOS method outperforms the other two models in the syntactic category and total accuracy even when they are trained with the double epochs. The Skip-gram ep10 method leads the semantic category.

6.3.2 Outlier Detection

We evaluated the models through the outlier detection framework proposed by (Camacho-Collados and Navigli, 2016). The results are shown in Table 4 below.

Model	OPP Score	Accuracy
CBOW ep10	**100.0**	**100.0**
CBOW ep5	**100.0**	**100.0**
Skip-gram ep10	99.414	95.312
Skip-gram ep5	99.414	95.312
CBOS ep5	99.609	98.437

Table 4: Outlier position percentage score and accuracy of the CBOS model and baselines on outlier detection task using the English Wikipedia Corpus for training.

The CBOW models reach a perfect score in both metrics regardless of epochs used to be trained, while our proposed CBOS model reaches almost a perfect score as well.

	English Wikipedia Corpus			Greek Wikipedia Corpus			Greek Web Content Corpus		
Model	Semantic	Syntactic	Total	Semantic	Syntactic	Total	Semantic	Syntactic	Total
CBOW ep10	40.21	**71.45**	50.47	32.71	45.16	39.81	21.01	55.26	43.16
CBOW ep5	35.19	71.11	46.99	25.14	42.93	35.29	20.03	54.42	42.27
Skip-gram ep10	**47.26**	63.80	**52.69**	**58.73**	41.73	49.03	**44.35**	51.07	48.69
Skip-gram ep5	43.19	61.68	49.26	51.79	42.88	46.71	41.27	52.27	48.38
CBOS ep5	42.94	68.08	51.20	52.72	**48.23**	**50.16**	41.16	**62.39**	**54.89**

Table 3: Accuracy of the CBOS model and baselines on word analogy task using three different datasets for training.

Model	ESSLI 2c	MEN	WS353	MTurk	Google	MSR	YP	Average
CBOW ep10	**0.622**	0.712	0.557	0.633	0.504	0.599	0.368	0.571
CBOW ep5	0.577	0.693	0.533	0.631	0.470	**0.601**	0.359	0.552
Skip-gram ep10	0.6	0.737	**0.703**	0.684	**0.521**	0.486	0.466	0.599
Skip-gram ep5	0.577	0.733	0.691	0.684	0.457	0.461	0.436	0.577
CBOS ep5	**0.622**	**0.743**	0.669	**0.701**	0.512	0.565	**0.475**	**0.612**

Table 5: Accuracy of the CBOS model and baselines on various word similarity benchmarks using the English Wikipedia Corpus for training.

	English Wikipedia Corpus			Greek Wikipedia Corpus			Greek Web Content Corpus		
Model	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
CBOW ep10	69.48	69.5	69.47	81.91	81.52	81.64	81.64	81.49	81.52
CBOW ep5	69.38	69.40	69.38	80.47	80.1	80.21	81.11	81.18	81.11
Skip-gram ep10	**70.46**	**70.49**	**70.47**	82.03	81.72	81.78	81.37	81.12	81.19
Skip-gram ep5	70.44	70.46	70.44	81.6	81.42	81.44	81.82	81.45	81.58
CBOS ep5	70.31	70.32	70.3	**82.46**	**82.10**	**82.18**	**83.37**	**83.07**	**83.17**

Table 6: Precision, Recall and F1 score of the CBOS model and baselines on Text Classification using Linear SVC algorithm and three different datasets for training.

6.3.3 Word Similarity

Table 5 shows the results of the models evaluated in a range of popular benchmarks used for word similarity. The framework used for this evaluation was developed by (Jastrzebski et al., 2017).

The proposed CBOS model outperforms the other two models in most benchmarks, as well as in average score. The Skip-gram model trained with double epochs achieves the highest score in two datasets (WS353, Google).

6.4 Extrinsic Evaluation

6.4.1 Text Classification

We divided each dataset in three parts: training, validation and test set. Every experiment was repeated 10 times and the result used is the mean value of Precision, Recall and F1 metrics in the test set. The algorithm used for the text classification task was Linear SVC with its default parameters.

The results in Table 6 show that the Skip-gram model achieves a slightly higher performance in the evaluation process for the English language.

Concerning the evaluation procedure for the Greek language, the CBOS model outperforms the other two models regardless of training epochs. In particular, the highest difference is achieved in Table 6 where our proposed model leads all metrics by 1 - 1.5 % .

6.4.2 Named Entity Recognition

For this task we used a bi-LSTM + CRF + chars embeddings model. We split the datasets in train, validation and test sets, use early-stopping, run every experiment 10 times with random seed and as result use mean value of F1 metric.

The results in Table 7 show that the CBOS model outperforms CBOW and Skip-gram models that have been trained with the same or double number of epochs on the English Wikipedia corpus.

Trained on the Greek Wikipedia Corpus, the CBOS model is between CBOW and Skip-gram models that have been trained with the same number of epochs. Concerning the Greek Web Content Corpus, the CBOS model slightly outperforms CBOW and greatly outperforms the Skip-gram model.

A useful observation is that using F1-val score, where early stopping is used, CBOS is best overall in both English and Greek Wikipedia datasets.

7 Conclusions

This paper aimed at producing high-quality word embeddings mainly for the Greek language, devising a new embedding method, the Continuous Bag-of-Skip-grams (CBOS). CBOS combines the benefits of the CBOW and Skip-gram approaches introduced in (Mikolov et al., 2013b). Because of its neat implementation, CBOS does not increase the computational cost of the training phase.

We presented in Section 6 that there are particular tasks where CBOW outperforms Skip-gram and vice versa. Since there is no one method that outperforms all others in all tasks, we strongly recommend the use of CBOS. It is evident that CBOS

	English Wikipedia Corpus		Greek Wikipedia Corpus		Greek Web Content Corpus	
Model	**F1-val**	**F1-test**	**F1-val**	**F1-test**	**F1-val**	**F1-test**
CBOW ep10	92.939	88.057	71.897	**72.335**	75.554	76.219
CBOW ep5	92.631	87.628	70.868	70.803	**75.792**	76.777
Skip-gram ep10	92.414	87.647	71.897	**72.335**	72.407	71.949
Skip-gram ep5	92.382	87.393	72.007	72.059	73.054	72.104
CBOS ep5	**93.055**	**88.977**	**72.220**	71.686	74.943	**77.145**

Table 7: F1 score in validation and test set of the CBOS model and baselines on Named Entity Recognition using three different datasets for training.

achieves a high performance in every task and the highest performance in most cases regardless of the training epochs.

Additionally, we believe that the CBOS method achieves higher performance in Greek language than English due to their different linguistics aspects (Lascaratou, 1998; Greenberg, 1963; Tzartzanos, 1963). For instance, Greek language tends to have longer and more complex sentences. Furthermore, the word order is one of the most important differences of the two languages. The English language is more structured, while the Greek language is more fluid. In conclusion, the CBOS method tends to learn more accurate word representations when the language is more complex and less structured.

The future work of this research could include a more extensive research on CBOS alternatives, which will be based on other criteria than the Greek word analogy task. For instance, the impact of the new embeddings used in an extrinsic task could be considered for comparison. This fine-tuning on the production of word embeddings could lead to significant improvements in extrinsic evaluation tasks.

Acknowledgments

We would like to thank Prof. Ion Androutsopoulos for the useful discussions and suggestions.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José Camacho-Collados and Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 43–50.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305.

Joseph Harold Greenberg. 1963. Universals of language.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Chryssoula Lascaratou. 1998. Basic characteristics of modern greek word order. *Constituent order in the languages of Europe*, 151:171.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North*

American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 196–203.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. 2019. Evaluation of greek word embeddings. *arXiv preprint arXiv:1904.04032*.

Stamatis Outsios, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis. 2018. Word embeddings from large-scale greek web content. *arXiv preprint arXiv:1810.06694*.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, 33(3-4):175–190.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Achilleas Tzartzanos. 1963. *Νεοελληνική σύνταξις* (modern greek syntax). B.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8.

Fei Wu and Daniel S Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 118–127. Association for Computational Linguistics.