THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 19, Number 2/2018, pp. 403–409

COMPUTING DISTRIBUTED REPRESENTATIONS OF WORDS USING THE COROLA CORPUS

Vasile Păiş, Dan Tufiş

”Mihai Drăgănescu” Institute for Artificial Intelligence, Calea ”13 Septembrie”, no. 13, 050711, Bucureşti,
Corresponding author: Vasile Păiş, E-mail: pvf2005@gmail.com

Abstract. We investigate the usability of the CoRoLa corpus for generating high quality vector representations of words for Romanian language. Different model parameters are tested and model quality is compared in three test cases: two word analogies data sets and a word similarity correlation with human judgment. Furthermore, we prove that CoRoLa provides superior word representations compared to other known Romanian corpora, such as the Wikipedia corpus.

Key words: neural networks, word embeddings, vector representation, CoRoLa.

1. INTRODUCTION

Distributed representation of words, also known as word embeddings, is a way of grouping together words with similar or related meaning. For this purpose, the words are represented in a vector space. Thus each word becomes a point inside an n-dimensional vector space. Therefore, the word embedding calculation can be considered as a mapping $V \rightarrow R^n : w \rightarrow \underline{w}$, where V is the vocabulary, and each word w from that vocabulary is mapped to a corresponding real-valued vector $\underline{w}$ of dimension n. The vector dimension is much smaller than the total number of words in the vocabulary.

The vast majority of modern natural language processing applications make use of the distributed representation of words because it provides two major advantages over other methods for word representation. The first advantage is the reduced size of input variables, deriving from the dimension used and not from the vocabulary size. The second and the most important advantage is the capture of word's meaning inside the representation.

A well known quotation by linguist John Rupert Firth states that "you shall know a word by the company it keeps" [1]. Making use of this statement, there are currently two main methods for automatically learning distributed representation of words from a large text corpus: Skip-gram and continuous bag of words (CBOW) [2]. These methods make use of artificial neural networks in order to automatically learn the word representations by building a language model.

From an historic point of view, Skip-gram and CBOW are a natural development following the initial model proposed by Bengio *et al*. [3]. In this work, a feedforward neural network with a linear projection layer and a non-linear hidden layer was used to learn the word vector representation and a statistical language model.

In the case of CBOW, as detailed in [2], a neural network is trained to predict the middle word given a context (a number of previous and next words, around the predicted word).

The Skip-gram model uses a similar approach, but instead of predicting one word based on the context, it uses a current word as input and tries to predict past and future words. Thus in a way it is a mirror architecture compared to CBOW. While in CBOW the context is used to predict one word, in the Skip-gram model one word is used to predict the context.

Distributed representation of words involving mapping one word to a vector poses a problem in the case of unknown words. These are words without a previously learned vector representation, also known as out-of-vocabulary (OOV) words. A solution to this problem is presented in the work of Bojanowski *et al*. [4].

In this paper, an extension to the Skip-gram model is introduced by calculating vector representations for character n-grams instead of actual words. The words are then represented as sums of the corresponding character n-grams forming the word. In this case unknown words can be represented using the same summation, performed when needed, after the model was trained.

2. REPRESENTATION OF WORDS IN THE COROLA CORPUS

The Reference Corpus for Contemporary Romanian Language (CoRoLa) [5] was constructed as a priority project of the Romanian Academy, between 2014 and 2017. It contains both written texts and oral recordings. Its aim was to cover major functional language styles (legal, scientific, journalistic, imaginative, memoirs, administrative), in four domains (arts and culture, nature, society, science) and in 71 sub-domains while taking into account intellectual property rights (IPR). With over 1 billion word tokens (written and spoken), CoRoLa (corola.racai.ro) is one of the largest fully IPR-cleared Reference Corpus in the world. CoRoLa is searchable via three interfaces (two for written part and one for the oral part) and it is supported but the KorAP corpus management platform, developed at Institute for German Language in Mannheim [6; 7].

Apart from the raw texts, the CoRoLa corpus contains linguistic annotations in the form of: phoneme, syllable, lemma, part of speech (POS) tagging, syntactic chunking, dependency parsing. Several of these annotation types were added using the TTL [8] service. However, in order to make fair comparison with word embeddings extracted for Romanian language [4], with the same application [https://fasttext.cc/docs/en/pretrained-vectors.html#models], we did not use any annotation, but only the wordforms.

Given the large amount of textual information available in the CoRoLa corpus, it becomes an obvious choice for training word representations using this corpus. The large number of words from different domains, means there are going to be lots of contexts capturing different word meanings. Furthermore, rare words are present in different domains. Therefore, compared to other Romanian corpora, such as the Romanian Wikipedia corpus, CoRoLa is presently the best choice for natural language processing tasks, including learning word representations.

The chosen algorithm for the distributed word representation model is the Skip-gram, introduced in [2], with the sub-word information added by character n-gram, as presented in [4].

The following model parameters were considered influential for the learned word representations: the size of the word vectors, minimum number of occurrences in the corpus for a particular word to be taken into consideration.

Dimensionality of the word vectors is usually considered important in capturing a more detailed meaning of a word and thus exploits context information present in large corpora. However, large corpora may contain misspelled words. Therefore the minimum occurrences parameter can be used to filter out such cases. We did experiments with several vector dimensions (100, 200, 300, 400, 500 and 600) and different occurrence thresholds (1, 5, 10, 20 and 50) as shown in Table 1.

3. RESULTS

The aim of this work was to produce a high quality vector representation of words using the word occurrences found in the CoRoLa corpus. For this purpose, it was necessary to have a way of comparing different results in order to understand which one is better. One way of doing this is to compare similar/related words indicated by a trained model with a known list of such words.

For example, a list of 10 similar words for the word “italia” (“italy”), produced by a word embedding representation is: “spania, franța, portugalia, olanda, belgia, grecia, anglia, milano, germania, ungaria”. This list contains other countries, which may be expected. However it also contains a city in Italy, “Milano”. If the similarity is meant as referring to “countryhood” then Milano is wrong. Yet, the obvious relatedness of Milano to Italy, might license it as a right answer. It’s not easy to say this is correct or wrong. Similarly with other words, it’s not easy to produce a list of all expected related words, especially taking into account the various possible meanings a word may have.

Instead of using lists, a similar approach to that used in Bojanowski *et al*. [4] is taken. The dataset used is the one described in Hassan and Mihalcea, 2009 [9], which is a manual translation of the WordSimilarity-353 dataset (also known as Finkelstein-353), as described in Finkelstein *et al*. [10]. This dataset contains a human similarity/relatedness judgement between pairs of words. The similarity/relatedness according to the word representation being learned is computed as a cosine distance between the corresponding word vectors. In order to compare the computed similarity with the human judgment, we calculated the Spearman's rank correlation coefficient [11] between the two. Results are presented in Table 1 for different model parameters.

In the work by Mikolov *et al*. [2] is also proposed a different kind of measure for the word representation accuracy. Instead of looking at single words, it seems better to look at the relation between pairs of words. For example we may consider the pairs "rege" - "bărbat" ("king" - "man") and "regina" - "femeie" ("queen" - "woman"). In this case we may ask the question: 'What is the word that is related to "woman" in the same way as "king" is related to "man" ?'. Using word representations, this question can be translated into:

vec("rege") - vec("bărbat") + vec("femeie")

The answer should be *vec("regină")*.

For the purpose of this work, we constructed automatically a list of questions and associated answers, using European capital cities and their corresponding countries. A total of 1892 questions and their expected answers were produced (SET1). Furthermore, we extracted a subset of questions using only words for which their number of occurrences is in top 30.000. This produced a number of 462 questions and their corresponding answers (SET2). These two sets are available online at [12].

Each trained model using different values for the parameters indicated above was tested against the questions and the first returned value was compared with the expected answer. A question was considered answered correctly only if the first produced similar word was exactly the same as the expected answer. The results are presented in Table 1. With a more relaxed evaluation (say Precision@3) the Accuracies on both sets are significantly improved[1]. Pre-trained vectors for the dimensions given in Table 1 are available for download at [12].

Table 1

Accuracy of different models by number of dimensions and minimum occurrences

Nr	Dimensions	Minimum Occurrences	Accuracy SET1	Accuracy SET2	Correlation with WS-353
1	100	1	20%	48%	51
2	100	5	26%	61%	53
3	100	10	25%	56%	51
4	100	20	26%	54%	49
5	100	50	22%	40%	49
6	200	1	23%	58%	54
7	200	5	31%	65%	52
8	200	10	31%	64%	51
9	200	20	35%	72%	49
10	200	50	35%	66%	50
11	300	1	20%	52%	54
12	300	5	31%	64%	51

[1] For instance, the 14th parameter combination in Table 1 gives 52% P@3 for SET1 and 92% for SET2.

(continued)

13	300	10	32%	67%	50
14	**300**	**20**	**35%**	**74%**	**52**
15	300	50	37%	72%	49
16	400	1	18%	48%	55
17	400	5	28%	61%	52
18	400	10	30%	64%	52
19	400	20	33%	64%	47
20	400	50	36%	72%	49
21	500	1	19 %	49 %	50
22	500	5	26 %	56 %	48
23	500	10	31 %	67 %	47
24	500	20	35%	69%	45
25	500	50	38 %	72 %	48
26	600	1	14 %	37 %	50
27	600	5	24 %	54 %	50
28	600	10	28 %	61 %	50
29	600	20	31 %	62 %	49
30	600	50	38 %	69 %	47

From the above table it can be seen that both number of dimensions and minimum number of occurrences plays an important role in the model's accuracy. The sizes of generated word embeddings are dependent on these two parameters, ranging from 9.5 Gb for 600 dimension vectors and no frequency threshold down to 133 Mb for 100 dimension vectors and 50 frequency threshold; the embeddings for the 14th parameter combination (300/20) in Table 1 need 632 Mb. CoRoLa is a large corpus and therefore an increased number of dimensions is beneficial for the word representation model, allowing more subtle meaning to be properly inferred. Yet, filtering the vocabulary, by using a frequency threshold and excluding infrequent words, most of which are typing errors, allows a further increase in the accuracy of the trained model, especially with regard to word analogies.

The authors of [4] published pre-trained vector representations of words for different languages, including Romanian, based on Wikipedia corpora. These are available at [13]. In Table 2 is presented a comparison for the accuracy of the pre-trained Romanian model based on Wikipedia and the model trained on CoRoLa, on the same number of dimensions.

Table 2

Comparison of pre-trained Wikipedia model with CoRoLa model

Model	Dimensions	Accuracy SET1	Accuracy SET2	Correlation with WS-353
Wikipedia	300	26%	63%	54
CoRoLa	300	35%	74%	52

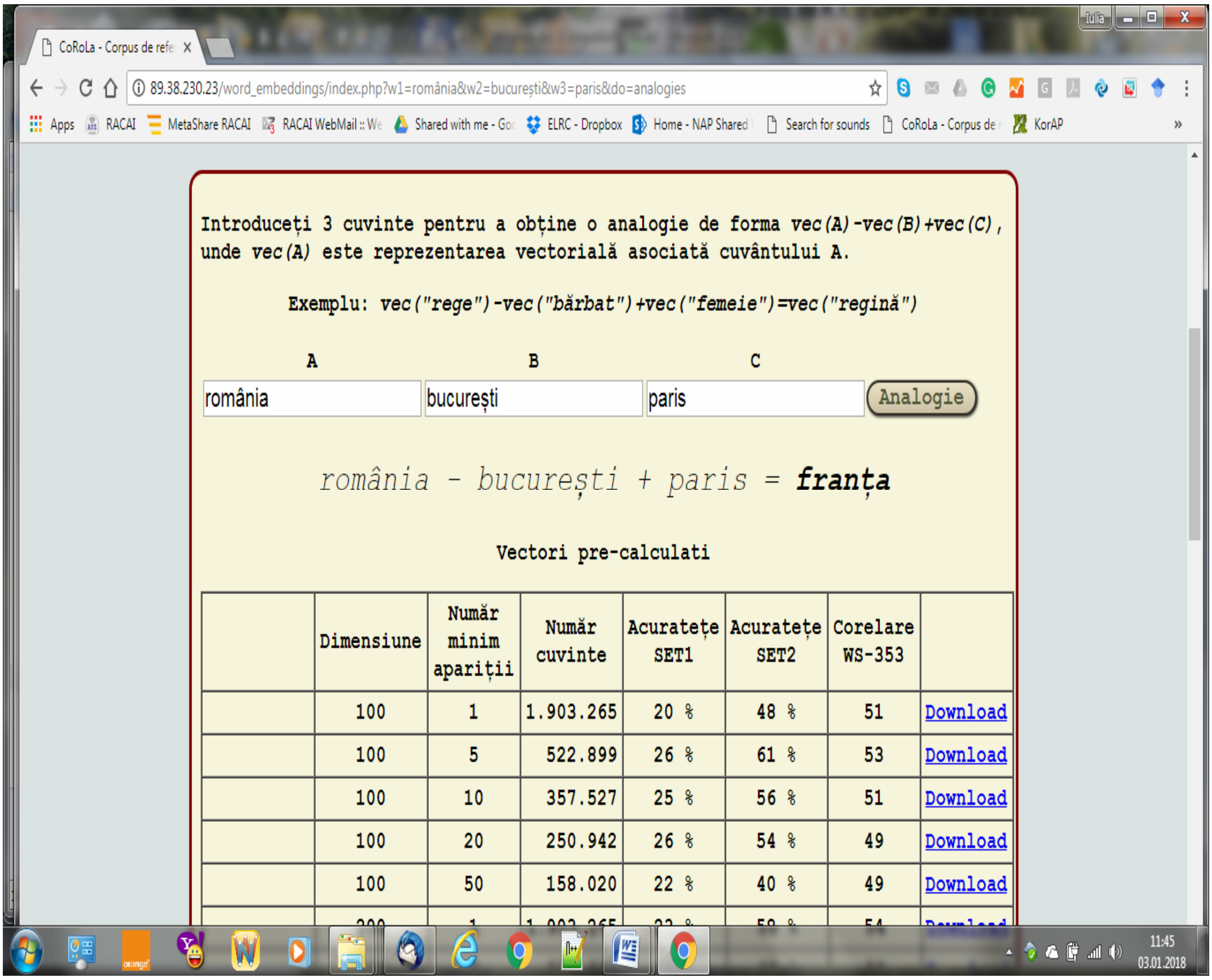
In Table 3 are given examples of 5 nearest neighbors for several words, obtained using the CoRoLa model. For each query only the first 5 values are shown in the table.

Table 3

Nearest neighbor examples obtained with the CoRoLa model

Word	spania	ilie	euro	sibiu	fizician
1	portugalia	dumitru	usd	brașov	biofizician
2	franța	stoica	dolari	cluj	astrofizician
3	italia	gheorghe	milioane	arad	fizicianul
4	grecia	valeriu	miliarde	sighișoara	matematician
5	olanda	florea	forinți	oradea	geofizician

In Fig. 1 it is shown a snapshot from the analogy interface available on-line. In Table 4 are exemplified several analogies obtained using the CoRoLa model. For each query only the first result is shown in Table 4.



89.38.230.23/word_embeddings/index.php?w1=românia&w2=bucurești&w3=paris&do=analogies

Introduceți 3 cuvinte pentru a obține o analogie de forma *vec(A)-vec(B)+vec(C)*, unde *vec(A)* este reprezentarea vectorială asociată cuvântului A.

Exemplu: *vec("rege")-vec("bărbat")+vec("femeie")=vec("regină")*

A: românia
B: bucurești
C: paris

românia - bucurești + paris = ***franța***

Vectori pre-calculati

	Dimensiune	Număr minim apariții	Număr cuvinte	Acuratețe SET1	Acuratețe SET2	Corelare WS-353	
	100	1	1.903.265	20 %	48 %	51	Download
	100	5	522.899	26 %	61 %	53	Download
	100	10	357.527	25 %	56 %	51	Download
	100	20	250.942	26 %	54 %	49	Download
	100	50	158.020	22 %	40 %	49	Download
	200	1	1.903.265	[illegible] %	58 %	54	Download

Fig. 1 – Analogy interface.

There are three words A, B, C and the model is used to compute *vec(A)–vec(B)+vec(C)*.

Table 4

Examples of analogies

A	B	C	A-B+C
roma	italia	franța	paris
tată	bărbat	femeie	mamă
regină	femeie	bărbat	rege
politician	cinstit	necinstit	politicianist
urât	frumos	alb	negru
frumos	urât	negru	alb
fizicianul	fizician	matematician	matematicianul
lingvistul	lingviști	informaticieni	informaticianul

Notice in the last two examples the morphology-based analogy (with the lemma-based vectors, such analogies wouldn't be available).

5. CONCLUSIONS AND FURTHER WORK

This work focused on obtaining efficient word representations for Romanian language using the CoRoLa corpus. Different model parameters were investigated in order to increase the accuracy of the trained model.

The results in Table 1 are consistent with those reported by Mikolov *et al.* in [2] for a large English corpus and those reported by Bojanowski *et al.* in [4]. This further confirms the representativeness of the CoRoLa corpus, as a Reference Corpus for Contemporary Romanian Language. The development of CoRoLa will continue within the DruKoLA [14] and EuReCo [15] projects led by the Institute for German Language (IDS) with activities on several directions: first, the detected errors will be remedied (typographical errors, missing diacritics, missing or wrong metadata, wrong annotations, etc.) then, new processed texts (written and oral) both monolingual and multilingual will be added and several exploitation facilities (user defined sub-corpora, comparative analytics, etc.). When the content of CoRoLa significantly changed, the word embeddings will be updated, possibly with other optimal parameters. We also plan to generate from CoRoLa other sets of vector representation, this time using also linguistic annotations (lemmas, POS, dependencies).

REFERENCES

1. J.R. Firth, *Papers in Linguistics 1934–1951*, London, Oxford University Press, 1957.
2. T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, 2013.
3. Y. Bengio, R. Ducharme, P. Vincent, *A neural probabilistic language model*, Journal of Machine Learning Research, **3**, pp.1137–1155, 2003.
4. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Enriching Word Vectors with Subword Information*, arXiv:1607.04606, 2016.
5. V. Barbu Mititelu, D. Tufiş, E. Irimia, *The Reference Corpus of Contemporary Romanian Language* (CoRoLa), Proceedings of the 11th Language Resources and Evaluation Conference – LREC'18, Miyazaki, Japan, European Language Resources Association (ELRA), 2018.
6. Bański, P., Diewald, N., Hanl, M., Kupietz, M., Witt, A., *Access Control by Query Rewriting. The Case of KorAP*, Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14), Reykjavik, European Language Resources Association (ELRA), 2014.
7. Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. and A. Witt, *KorAP Architecture – Diving in the Deep Sea of Corpus Data*, Calzolari, Nicoletta *et al.* (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, European Language Resources Association (ELRA), 2016.

8. D. Tufiş, R. Ion, A. Ceauşu, D. Ştefănescu, *RACAI's Linguistic Web Services*, Proceedings of the 6th Language Resources and Evaluation Conference – LREC'08, Marrakech, Morocco, 2008.

9. S. Hassan, R. Mihalcea, *Cross-lingual semantic relatedness using encyclopedic knowledge*, Proc. EMNLP, 2009.

10. L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, *Placing search in context: the concept revisited*, WWW, pp. 406–414, 2001.

11. C. Spearman, *The proof and measurement of association between two things*, The American Journal of Psychology, **15**, *1*, pp. 72–101, 1904.

12. Pre-trained CoRoLa word vectors, analogies application and data sets: http://89.38.230.23/word_embeddings/

13. Pre-trained word vectors using Wikipedia corpus, https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

14. Cosma, R., Cristea, D., Kupietz, M., Tufiş, D., Witt, A., *DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora*, Proceedings of the fourth CLMC, LREC 2016, Portoroz, Slovenia, 28 May 2016, European Language Resources Association (ELRA).

15. Kupietz, M., Witt, A., Bański, P., Tufiş, D., Cristea, D., Váradi, T., *EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research*, Bański, P, Kupietz, M., Lüngen, H., Rayson, P., Biber, H., Breiteneder, E., Clematide, S., Mariani, J., Stevenson, M,/Sick, T. (eds.), Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017, Mannheim, Institut für Deutsche Sprache, pp. 15–19.

Received January 5, 2018