Russian Named Entities Recognition and Classification Using Distributed Word and Phrase Representations

Roman Ivanitskiy
ITMO University
Saint-Petersburg
Russia
litemn@yandex.ru

Alexander Shipilo
Saint-Petersburg State University
Saint-Petersburg
Russia
ITMO University
Saint-Petersburg
Russia
alexandershipilo@gmail.com

Liubov Kovriguina
ITMO University
Saint-Petersburg
Russia
lyukovriguina@corp.ifmo.ru

Abstract

The paper presents results on Russian named entities classification and equivalent named entities retrieval using word and phrase representations. It is shown that a word or an expression's context vector is an efficient feature to be used for predicting the type of a named entity. Distributed word representations are now claimed (and on a reasonable basis) to be one of the most promising distributional semantics models. In the described experiment on retrieving similar named entities the results go further than retrieving named entities of the same type or named entities-individuals of the same class: it is shown that equivalent variants of a named entity can be extracted. This result contributes to the task of unsupervised entities and semantic relations clustering and can be used for paraphrase search and automatic ontology population. The models were trained with word2vec on the Russian segment of parallel corpora used for statistical machine translation. Vector representations were constructed and evaluated for words, lexemes and noun phrases.

1 Introduction

Model of distributed word and phrase representations introduced by Mikolov in 2013 (Mikolov et al., 2013) has proved its efficiency on a variety of languages and tasks in natural language processing and got a number of extensions since its appearance. It provides a faster and more accurate implementation of the models relying on the basic idea of distributional semantics known as "similar words occur in similar contexts". Mikolov et al. have shown that "word representations computed using neural networks are very interesting because the learned vectors explicitly encode many linguistic regularities and patterns, and ... many of these patterns can be represented as linear translations" (Mikolov et al., 2013). This paper presents the results of word2vec[1] application to the traditional NLP task - named entity recognition (NER) - for the Russian language. Results concerning NER classification can contribute to the pool of evaluation data and extend existing distributional semantic models for Russian, i.e., RusVectores[2].

NER recognition and classification can be successfully done using a large number of techniques and resources, especially technologies of Semantic Web and knowledge bases like DBPedia[3], which provides semantic search over billions of entities. DBpedia Spotlight[4], a tool for automatically annotating mentions of DBpedia resources in the text, can skip the problem of NER annotation for newswire corpora, nonfiction corpora, datasets of medical records, etc. However, some genres of human discourse produce texts that lack such resources and demand considerable efforts on its annotation: spoken language gives a plenty of examples of occasional abbreviations, unpredictable names distortion of personalia, toponyms and organizations. Moreover, there has emerged a recent activity on paraphrase search. This determined the interest to analyze the response of the trained word2vec model given a named entity as a stimulus. Before applying word2vec to spoken corpora we decided to test its ability to cluster named entities with the same label and extract semantic equivalents for a given named entity on Russian segment of parallel corpora used for ma-

[1] Word2vec is a group of models (and software) for unsupervised word representations learning.

[2] Cf. http://ling.go.mail.ru/dsm/en/

[3] Cf. http://wiki.dbpedia.org/

[4] Cf. http://spotlight.dbpedia.org/

chine translation. Two experiments are described in the paper. The first one learns SVM classifier on the FactRuEval[5] training dataset, the second experiment analyses lists of entities with the highest value of the cosine measure with the named entity-stimulus. Both experiments are done on 4 training models: models 1 and 2 were trained on a 1 billion corpus (word forms and lexemes respectively) and models 3 and 4 were trained on a 100 million corpus (a subset of the larger) which has been annotated with noun phrases to extend word representations to noun phrase representations.

2 Related Work

There exists a considerable number of studies on NER on English texts evaluating various types of algorithms, but Russian NER has been mostly done using rule-based algorithms and pattern matching whereas recent studies focus on words embeddings as a feature for training NER classifiers (Turian et al., 2010), on news corpora (Siencnik, 2015), (Seok et al., 2016), microblog posts (Godin et al., 2014), (Kisa and Karagoz, 2015), CoNLL 2003 Shared Task Corpus and Wikipedia articles.

Segura-Bedmar et al. (Segura-Bedmar et al., 2015) describe a machine learning approach that uses word embedding features to recognize drug names from biomedical texts. They trained the Word2vec tool on two different corpora: Wikipedia and MedLine aimed to study the effectiveness of using word embeddings as features to improve performance of the NER system. To evaluate approach and compare it with previous work, they made a series of experiments on the dataset of SemEval-2013 Task 9.1 Drug Name Recognition. Demir and Ozgur (Demir and Ozgur, 2014) developed a fast unsupervised method for learning continuous vector representations of words, and used these representations along with language independent features to develop a NER system. They evaluated system for the highly inflectional Turkish and Czech languages. Turkish datasets contained 63.72M sentences that correspond to a total of 1.02B words and 1.36M hapax legomena. Publicly available data crawled from Czech news sites provided by the ACL machine translation workshop were used for the Czech language. This dataset contained 36.42M sentences corresponding to 635.99M words and 906K hapax legomena.

A number of papers describe experiments that go beyond word representations and "construct phrase embeddings by learning how to compose word embeddings using features that capture phrase structure and context" (Yu and Dredze, 2015), (Lopyrev, 2014). However, "phrase" notion in these works is quite vague and varies considerably. Yin and Schultze stress that "generalized phrases ... include conventional linguistic phrases as well as skip-bigrams. ... Socher et al. use the term "word sequence". Mikolov et al. use the term "phrase" for word sequences that are mostly frequent continuous collocations" (Yin and Schütze, 2014). For the purposes of the described experiment accurate noun phrase extraction is crucial, because items of the noun phrase can be rare words but the whole phrase can occur in frequent contexts (about processing rare words in distributed word representations models see paper(Guthrie et al., 2006)).

3 Data Preparation

3.1 Datasets

Four datasets were built to train distributed word representations on the basis of FactRuEval training dataset and Russian parts of parallel corpora used to train statistical machine translation systems[6]. The list of all used corpora is given below:

- Russian subcorpus of Multilingual UN Parallel Text 2000—2009,
- Europarl,
- News,
- FactRuEval,
- Russian subcorpus of Yandex parallel corpus,
- Russian subcorpus of Czech-English-Russian parallel corpus.

Total size of these corpora is 1 billion tokens.

Datasets will be from now on referred to as Dataset 1, Dataset 2, Dataset 3 and Dataset 4. They were used to train word2vec models with the same indices. Basic preprocessing included removal of xml/html tagging, timestamps and URLs.

[5]Cf. http://github.com/dialogue-evaluation/factRuEval-2016

[6]Cf. http://www.statmt.org/

Dataset 1. This corpus is built of wordforms of 1 billion corpora and has no linguistic preprocessing except tokenization. Training entity is word form.

Dataset 2. This is 1 lemmatized billion corpus. Tagging was performed using Mystem morphological analyzer[7] supporting homonymy resolution. Training entity is lexeme.

Dataset 3. This is 100 million subcorpus of the above corpus. Training entities are wordforms and noun phrases.

Dataset 4. This is lemmatized 100 million subcorpus of the above corpus. Training entities are lexemes and noun phrases (also represented by lexemes).

3.2 Noun Phrase Extraction for Corpora 3 and 4

For the given task, a noun phrase may include more than one named entity, therefore, to provide equal context probability smaller noun phrases were extracted from the complex ones (i.e string "Government of Krasnoyarsk Krai" (label:organization) is represented by the whole noun phrase and its smaller part: noun phrase "Krasnoyarsk Krai" (label:location). For these cases sentences are duplicated in the corpus for each embedded noun phrase. Noun phrases are extracted using the following procedure:

- input sentences are tokenized, tagged and parsed using SemSin syntactic parser that produces a labelled syntactic tree for the input sentence(Kanevsky and Boyarsky, 2012);
- the NP extraction algorithm finds all word sequences depending from every noun within the sentence and writes these sequences as a candidate noun phrase;
- candidate noun phrases that contain no symbols in uppercase are filtered out.

4 Evaluation Procedure

System performance was evaluated using the above mentioned manually tagged FactRuEval test dataset. It has 3 basic types of named entities: name of persons, organizations and locations. For the first experiment a string containing named entity was sent to classifier and it produced its label. For datasets 1 and 2 evaluation dataset was cut to named entities represented by single word forms/lexemes, datasets 3 and 4 were evaluated on the whole test set (see results in Tables 2–5 of Section 6). For the second experiment named entities from the training FactRuEval dataset were used as stimuli. For datasets 1 and 2 the stimuli list included only unigrams and for datasets 3 and 4 the list was built of 20% unigrams and 80% of noun phrases length from 2 to 5. Each stimulus was fed to the trained word2vec model that generated a response list of 10 NE-candidates having highest cosine measures. Candidate NEs were manually tagged as true if a candidate was a named entity and had the same class as the stimulus, and false otherwise. Evaluation results are presented in Table 6, section 6.

5 Experiment Setup

The overall architecture of the system can be seen in Fig. 1. Software used includes open source word2vec toolkit[8], Java libraries for word2vec[9], Weka[10] and NLP software mentioned in Section 3.

Both experiments workflow comprises the following steps:

1. Data collection and cleansing;
2. Data linguistic processing (tokenization, sentence segmentation, tagging, parsing);
3. NP extraction;
4. Model training and evaluation on wordforms (trained model 1);
5. Model training on evaluation on lexemes (trained model 2);
6. Model training and evaluation on noun phrases (trained model 3 and 4);
7. Building stimuli lists for each model;
8. Experiment 1 on NE classification;

[7] Cf. `https://tech.yandex.ru/mystem/`

[8] Cf. `https://code.google.com/archive/p/word2vec`

[9] Cf. `http://deeplearning4j.org/`

[10] Cf. `http://www.cs.waikato.ac.nz/ml/weka/`

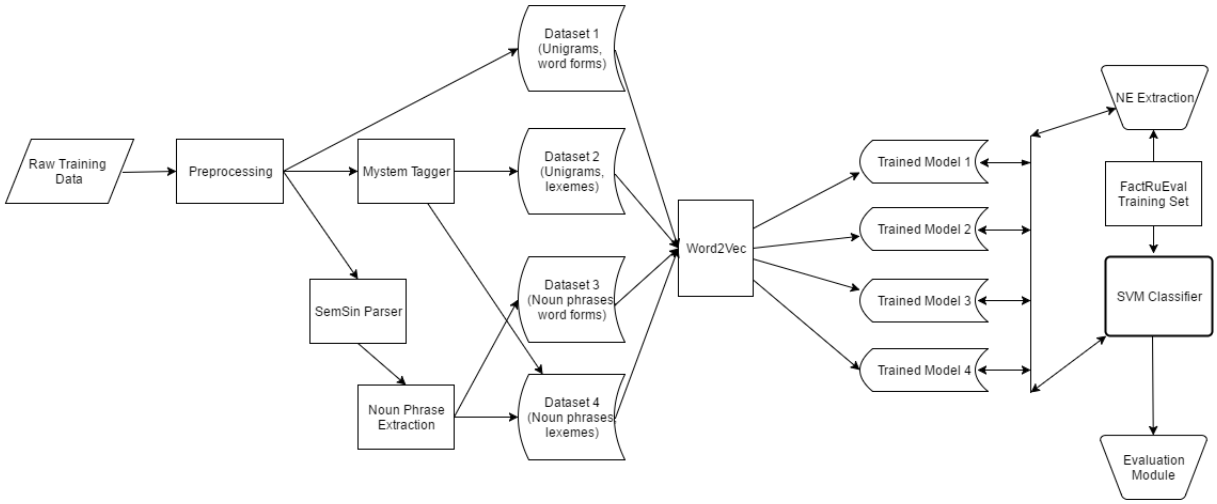

Figure 1: Workflow of Named Entity Recognition System Using Distributed Word and Phrase Representations

9. Experiment 2 on NE prediction and classification;

10. Evaluation.

Experiment 1 detailed plan. SVM classifier was learned on FactRuEval training set. NE word2vec vectors were used as feature vectors (dimension was set to 200). FactRuEval test set was used to test the classifier that is sent a NE-unigram or a NE-noun phrase and returns its label.

Experiment 2 detailed plan. Unigrams and noun phrases from the stimuli lists were sent to the trained word2vec models. Each model returned a list of 10-best candidates for each stimulus that included both words and phrases (for models 3 and 4). Percent of named entities having the same label as the stimulus was count.

6 Results and Discussion

Experiment 1: NE Label Prediction Evaluated on FactRuEval Training and Test Datasets

Figures 2–5 below show output of SVM classifier after dimensionality reduction using t-SNE algorithm[11] for all 4 training models. Distribution of NE labels conforms with the well-known fact that in many cases it is difficult or impossible to distinguish organizations and locations[12]. Classification quality was evaluated with f-score measure, results are given in tables 2–5. The system shows competitive quality in comparison to other machine learning or rule-based algorithms developed for the Russian language according to the report provided by the FactRuEval committee in 2016 (Starostin et al., 2016), see Table 1. In Table 1 minimum and maximum values for precision, recall and f-score are given. Average values for the performance of 13 NER systems that took part in the competition are given in round brackets. If we compare state-of-the-art performance with the performance of the described system (for model 4), based on distributed word representations approach, we can see that the system shows average results for locations (0.86 f-score) and persons (0.89 f-score) and outperforms state-of-the-art systems in retrieving organizations (0.79 vs 0.68 f-score). NE-unigrams are classified with very high f-scores (0.99, 0.96 and 0.97 f-scores for persons, locations and organizations respectively acc. to model 2). It can be seen from figure 3 that points corresponding to three NE types interfere less showing better classification results. This is a common feature for models 3 and 4 that both were trained on datasets containing lemmas, whereas models 1 and 3 (see fig. 2 and 4) were trained on datasets with wordforms and the areas corresponding to each NE are very vague. Persons names are classified with the highest f-score in all 4 models that is quite predictable, because sometimes distinguishing between locations and organizations is a non-trivial task (i.e. sometimes it can not be made clear from the context what is mentioned - a social institute (organization) or a building it occupies (location). Both for NE-single words and

[11] Cf. `https://lvdmaaten.github.io/tsne/`

[12] In Figures 2–5, 0 corresponds to organizations, 1 - to locations, 2 - names of persons.

NE-phrases results show importance of lemmatization before computing word embeddings for the inflectional languages with rich morphology, like Russian, even when a large corpus is used.

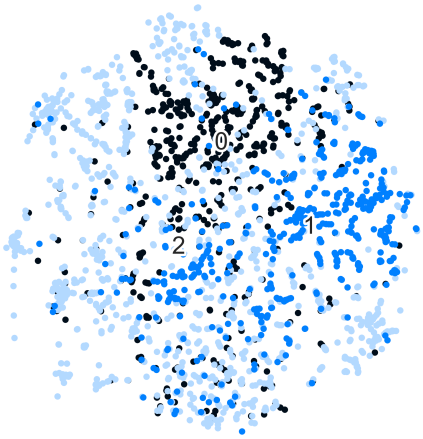

Figure 2: SVM Class Distribution for Model 1

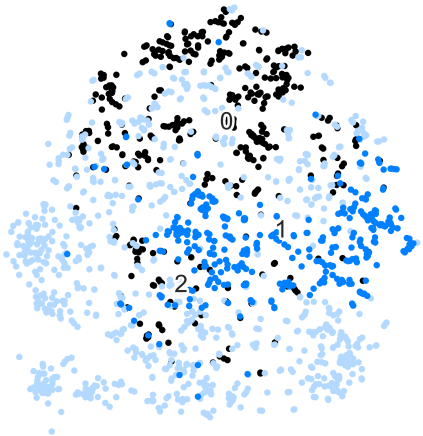

Figure 3: SVM Class Distribution for Model 2

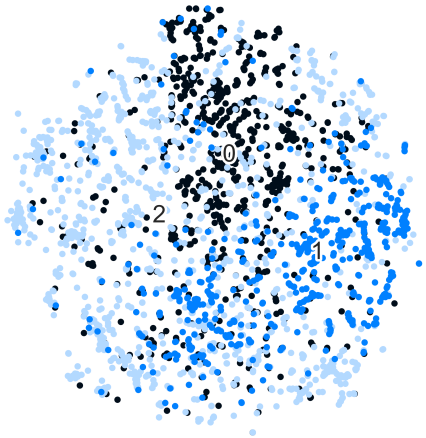

Figure 4: SVM Class Distribution for Model 3

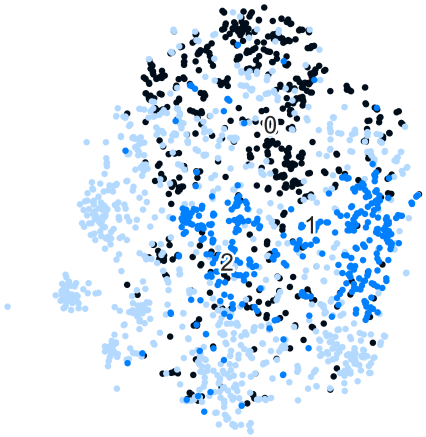

Figure 5: SVM Class Distribution for Model 4

Experiment 2: NE Search and Classification Using Word2vec Semantic Similarity Vectors

Quality was evaluated with f-score measure, percent of true positives is given in Table 5. The overall quality is not high, still it is possible to find and predict the class of unlabelled named entities which vectors have high cosine measure with the vector of the labelled NE. Trained model 1 produces high f-score values due to evaluation limitations: in models 1 and 2 only unigrams are considered. Consequent comparison of trained models 2-4 confirms that quality improves when noun phrases are predicted.

Given a word or a phrase, word2vec is capable to retrieve linguistic units that are involved in some semantic relation with the given one: synonyms, items of the same paradigmatic class, associations. But what can be found in the semantic similarity space of a named entity? In this experiment it is assumed that among words and phrases

NE label	Precision	Recall	F-score
Location	0.89-0.96 (0.93)	0.68-0.86 (0.80)	0.76-0.91 (0.86)
Person	0.90-0.96 (0.93)	0.73–0.92 (0.86)	0.81-0.93 (0.89)
Organization	0.74–0.87 (0.80)	0.26–0.76 (0.61)	0.39-0.79 (0.68)

Table 1: State-of-the-art performance of NER systems for the Russian language.

NE label	Precision	Recall	F-score
Location	0.93	0.79	0.85
Person	0.83	0.74	0.89
Organization	0.81	0.94	0.77

Table 2: Classification Accuracy for Model 1.

NE label	Precision	Recall	F-score
Location	0.96	0.96	0.96
Person	0.99	0.99	0.99
Organization	0.96	0.98	0.97

Table 3: Classification Accuracy for Model 2.

NE label	Precision	Recall	F-score
Location	0.87	0.77	0.81
Person	0.80	0.88	0.84
Organization	0.75	0.72	0.73

Table 4: Classification Accuracy for Model 3.

NE label	Precision	Recall	F-score
Location	0.88	0.84	0.86
Person	0.90	0.87	0.89
Organization	0.86	0.83	0.79

Table 5: Classification Accuracy for Model 4.

which vectors have high cosine measure with the vector of a named entity equivalent names of a named entity can be found. This turned out to be true for 48% of organizations, 50% of locations, 57% of person names (acc.to model 4). In 30% of cases more than 3 equvalent names are found among first 10 responses to the NE-stimulus. Be-

Model	Location	Person	Organization
Model 1	81,29	54,03	64,61
Model 2	62,17	52,13	46,90
Model 3	55,43	57,84	46,26
Model 4	67,63	68,21	49,45

Table 6: Unlabelled NE Prediction Accuracy on Distributed Representations.

low some examples are provided, only English translations are given. NE-stimulus is the first item in the list, given in italics, the rest items are responses. Equivalents (that can be paraphrases or alternative names) are given in bold.

- *The Prosecutor General:* **ATTY GEN**, **ATTY GEN of Russia**, **RF ATTY GEN**, Deputy Prosecutor General, **RF Prosecutor General**, **RF Prosecutor**, General Prosecutor Office, **Prosecutor General of Russia**, Prosecutor General of Ukraine, Prosecutor General of Moscow

- *Latin America:* Latin, South America, **Counties of Latin America**, **Latin American countries**, South-East Asia, **Countries of South America**, China, Country, Eastern Europe

In most cases, the list of responses contains individuals of the same class as the stimulus: i.e. given the name of a region in Russia, it will return a list of other Russian regions. Among the NE-candidates for the city stimuli wrongly lemmatized city names and toponyms misspellings were found, which can be also used to eliminate lemmatization or spelling mistakes.

7 Future Work

Future work implies development of a stable and comprehensive model of distributed noun phrase representations that will extend existing resources for the Russian language. Admissible results on NE prediction using response word2vec lists allow to continue with the experiments on NE recognition from noisy texts and spoken language. Ability of distributed word representations to capture paraphrases and lexical variants of named entities can be used in algorithms of paraphrase search and similar entities and events clustering.

8 Acknowledgements

This work was partially financially supported by the Government of the Russian Federation, Grant 074-U01.

References

Hakan Demir and Arzucan Ozgur. 2014. Improving Named Entity Recognition for Morphologically Rich Languages Using Word Embeddings. In *13th International Conference on Machine Learning and Applications, ICMLA 2014, Detroit, MI, USA, December 3-6, 2014*, pages 117–122.

Frederic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2014. ACL W-NUT NER shared task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A Closer Look at Skip-gram Modelling.

Evgeniy Kanevsky and Kirill Boyarsky. 2012. The Semantic-and-Syntactic Parser SEMSIN. In *International conference on computational linguistics Dialog-2012 (-2012)*, Bekasovo, Russia.

Kezban Dilek Kisa and Pinar Karagoz. 2015. Named Entity Recognition from Scratch on Social Media. In *Proceedings of the 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE 2015) co-located with the 26th European Conference on Machine Learning / 19th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2015), Porto, Portugal, September 7, 2015.*, pages 2–17.

Konstantin Lopyrev. 2014. Learning Distributed Representations of Phrases.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Isabel Segura-Bedmar, Vıctor Suarez-Paniagua, and Paloma Martınez. 2015. Exploring Word Embedding for Drug Name Recognition. In *13th International Conference on Machine Learning and Applications, ICMLA 2014, Detroit, MI, USA, December 3-6, 2014*, pages 117–122.

Miran Seok, Hye-Jeong Song, Chan-Young Park, Jong-Dae Kim, and Yu seop Kim. 2016. Named Entity Recognition using Word Embedding as a Feature. *International Journal of Software Engineering and Its Applications*, 10(2).

Scharolta Katharina Siencnik. 2015. Adapting Word2vec to Named Entity Recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243.

A. Starostin, Bocharov V., Alexeeva S., and Bodrova A. 2016. FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow, June 1–4, 2016*.

Joseph Turian, Lev-Arie 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, Uppsala, Sweden, July. Association for Computational Linguistics.

Wenpeng Yin and Hinrich Schütze. 2014. An Exploration of Embeddings for Generalized Phrases. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Student Research Workshop*, pages 41–47.

Mo Yu and Mark Dredze. 2015. Learning Composition Models for Phrase Embeddings. *TACL*, 3:227–242.