Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

IMPROVING DISTRIBUTIONAL SEMANTIC MODELS USING ANAPHORA RESOLUTION DURING LINGUISTIC PREPROCESSING

Koslowa O. (evezhier@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Norway

In natural language processing, distributional semantic models are known as an efficient data driven approach to word and text representation, which allows computing meaning directly from large text corpora into word embeddings in a vector space. This paper addresses the role of linguistic preprocessing in enhancing performance of distributional models, and particularly studies pronominal anaphora resolution as a way to exploit more co-occurrence data without directly increasing the size of the training corpus.

We replace three different types of anaphoric pronouns with their antecedents in the training corpus and evaluate the extent to which this affects the performance of the resulting models in lexical similarity tasks. CBOW and SkipGram distributed models trained on Russian National Corpus are in the focus of our research, although the results are potentially applicable to other distributional semantic frameworks and languages as well. The trained models are evaluated against RUSSE'15 and SimLex-999 gold standard data sets. As a result, we find that models trained on corpora with pronominal anaphora resolved perform significantly better than their counterparts trained on baseline corpora.

Keywords: anaphora resolution, distributional semantics, word2vec, semantic similarity, vector space models, neural embeddings

РАЗРЕШЕНИЕ АНАФОРЫ В ОБУЧАЮЩЕМ КОРПУСЕ КАК СПОСОБ УЛУЧШЕНИЯ КАЧЕСТВА ДИСТРИБУТИВНО-СЕМАНТИЧЕСКИХ МОДЕЛЕЙ

Козлова О. (evezhier@gmail.com)

Национальный Исследовательский Университет
Высшая Школа Экономики, Москва, Россия

Кутузов А. (andreku@ifi.uio.no)

Университет Осло, Норвегия

Ключевые слова: разрешение анафоры, дистрибутивная семантика, word2vec, семантическая близость, векторные репрезентации лексики, искусственные нейронные сети

1. Introduction

Natural language semantics is the level of human language which is least formalized and understood: there is no general agreement even on what exactly to consider the meaning of a language unit. Thus, modeling it computationally is a very ambitious task. Distributional semantics is based on the hypothesis expressed in [Firth 1957] that meaning is composed of typical contexts in which a given unit occurs. This conceptualization of meaning can be represented with a vector space induced from large and representative corpora (with contexts as dimensions and frequencies as numeric values of components), so semantics becomes computationally tractable.

This approach to meaning representation has been studied for several decades. However, important events happened in 2013, when prediction-based models became popular, particularly Continuous Bag-of-Words (CBOW) and Continuous SkipGram algorithms proposed in [Mikolov et al. 2013] and implemented in *word2vec* software tool. Dense vectors generated by such models are called 'embeddings', and they are induced using machine learning techniques, such as artificial neural networks.

These algorithms were able to dramatically reduce computational complexity of training vector semantic models, without compromising performance [Baroni et al. 2014]. Additionally, they showed some interesting properties, such as geometrical vector operations reflecting semantic relations. This quickly resulted in these algorithms becoming an established standard in the field, used in wide spectrum of natural language processing applications. They are employed in constructing semantic maps of language and high-level processing, such as representing phrase and text semantics. This in turn forms a basis for various practical tasks: sentiment analysis, machine translation, information retrieval, fact extraction and natural language generation.

In terms of neural word embeddings, researchers traditionally search for ways of performance improvement in polishing learning algorithms and optimizing hyperparameters. However, in the presented research we deal with linguistic preprocessing of training corpora. Particularly, we describe an experiment of merging achievements of computational discourse analysis and distributional semantics. We address linguistic phenomenon of anaphora and treat its automatic resolution as a means to disclose 'under-the surface' word contexts to supply more training material for distributional models.

The experiment described further deals with three types of pronominal anaphors: relatives, reflexives and personal pronouns. As a rule, when training a distributional model, these entities are considered to be stop words and discarded during corpora preprocessing. We argue, instead, that this leads to waste of training resources, and models would benefit from associating proforms (elements with dependent semantic interpretation) with their antecedents. We show that models trained on Russian National Corpus [Plungian 2005] (further RNC) with resolved pronominal anaphora perform significantly better in standard lexical similarity tasks.

The structure of the paper is as follows: first, we present a brief overview of known approaches to linguistic preprocessing in distributional semantics, clarify related terminology and account for the state-of-the-art anaphora resolution for Russian. Next, we describe setting of our experiment, tools and evaluation methods. Then we discuss experimental results and outline the directions for the future work.

2. Related work

The topic of linguistic preprocessing of training corpora for distributional models is not novel. In [Pekar, 2004] some possible tips (lemmatization, morphological and syntactic analysis, rare context words removal) are explored for English. Lemmatization is considered most valuable, as apart from increasing model performance, it also reduces feature space. As stated in [Baroni, 2008], the preprocessing pipeline largely depends on availability of resources for a particular language and their quality, which may be an obstacle for advanced linguistic analysis. Tokenization, lemmatization and POS-tagging are considered to be baseline stages.

As for the connection between distributional semantics and discourse analysis, applying vector models to anaphora and coreference resolution is a research topic which draws an increasing amount of attention. Competitive results have already been demonstrated, as in [Clark, 2015]. However, to the best of our knowledge, the extent to which anaphoric associations could enrich the quality of the models themselves has been largely unexplored both for English and Russian.

The most relevant experiment to the one described in the present paper involved coreference chains mined from a large corpus as a training resource for a model used for semantic tasks [Adel, Schütze 2014]. This resulted in better precision in detecting antonyms. The authors acknowledge coreference chains to be valuable as a supplementary resource for creating semantic representations. However, it is reported that coreference chains alone encompass "only a small subset of word-word relations encoded in raw text". Our workflow addresses this issue.

3. Linguistic preprocessing of training corpora

Although it is possible to construct representations directly from tokenized text corpora, natural language texts are full of complexities, which may eventually cause significant adverse effect on the results. Considering this, performing minimal linguistic preprocessing is standard practice before training a distributional model. Constituent stages depend on many factors including the nature of the primary task, and may be very specific. However, there exists a number of common procedures implemented before the model actually starts learning vectors for words.

3.1. Tokenization

To model meaning through distribution, one first determines semantics of which language units is of interest. Thus, token segmentation is the only preparatory step required by vector models themselves. Minimal additional preprocessing starts when tokens are converted to lowercase, which is not obligatory but improves performance by eliminating the difference between capitalized and non-capitalized word forms. Note that some well-known publicly available models for English (for example, Google News model released along with *word2vec* code) lack even this basic preprocessing.

3.2. Morphological analysis: lemmatization and PoS-tagging

Replacing word tokens with their lemmas (normal forms or word types) has a twofold purpose: to reduce vocabulary size and to 'squeeze' all word forms co-occurrence information into one vocabulary item. This is especially important for rich morphology languages like Russian, where each noun may have at least 12 word forms, and having a separate vector for each of them is usually impractical (except when one needs to exploit relations between different grammatical forms of the same word). This stage involves specific linguistic tools, since sloppy analysis may even decrease vector quality. Sometimes lemmas are additionally supplied with part-of-speech tags, which helps to resolve ambiguity where lemmatization alone does not suffice [Kutuzov, Andreev 2015]. For example, Russian word "*печь*" can be interpreted either as a noun (*stove*) or as a verb (*to bake*). Without PoS information, a model would try to learn one vector for both words, which obviously does not make sense. At the same time, adding PoS tags makes it two words: "*печь_S*" and "*печь_V*", which then acquire two different vector representations.

3.3. Stop words

Some lexical units are considered to be not very useful in semantic tasks, and thus are filtered out from the corpus before training to get rid of unwanted 'noise'[1]. Such stop words are divided into two large groups:

1. **Functional words** (prepositions, numbers, conjunctions, etc). They do not possess their own meaning, and thus it does not make sense to spend training time on them.
2. In some tasks, it is useful to filter out **very frequent** (maybe, domain-specific) words or to downsample them during training so that their influence were limited.

Pronouns are as a rule considered to be examples of the first stop words type. Note that in the presented research we do not remove them from the training corpus, but replace them with their antecedents. In the next section we describe the role of pronominal anaphora in training corpora and give an account of anaphora resolution systems for Russian.

4. Pronominal anaphora and its resolution for Russian

Anaphora is a linguistic phenomenon whereby the interpretation of an occurrence of one expression depends on the interpretation of an occurrence of another or whereby an occurrence of an expression has its referent supplied by an occurrence of some other expression in the same or another sentence [King, Jeffrey C. 2013]. It is closely associated and sometimes confused with coreference. In both cases two or more expressions have the same referent, but the latter is much broader and does not presuppose interpretation dependency. For instance:

(1) a) *Мальчик$_i$, который$_i$ сидел за столом, был задумчив.*
The boy$_i$ who$_i$ was sitting at the table was lost in thought.
b) *Вася$_i$ сидел за столом. Мальчик$_i$ был задумчив.*
Vasya$_i$ was sitting at the table. The boy$_i$ was lost in thought

a) is the case of anaphora, since it is not possible to assign semantic interpretation to *'который'* ('who') without considering its co-indexical *'Мальчик'* ('The boy'). The unit which is dependent in its interpretation is called the **anaphor**, while the one which provides interpretation is the **antecedent**.
b) is the case of so-called coreferential noun phrases, each co-indexed unit possesses its own lexical meaning, but they all refer to the same person.

In this research, we limit ourselves to anaphora, leaving coreference for future work.

---

[1] Of course, it depends on the nature of the task. For example, in authorship attribution removing stop words can be undesirable.

Anaphoric pronouns occupy top positions in frequency lists for most languages. Russian is no exception. Table 1 provides frequency statistics for 3 pronominal anaphora types we enumerated earlier, in 2 well-known Russian corpora after PoS-tagging.

Table 1. Frequencies of anaphoric pronouns (instances per million)

Corpus	Relatives	Reflexives	Personal pronouns
RNC[2]	4,264.8	6,486.2	31,377.7
Open Corpora[3]	5,891.4	5,367.3	18,961.0

It is obvious that these pronouns are in fact coreferential representations of meaningful words. They convey additional information on the distribution of the latter in a covert, non-explicit way.

Our major hypothesis is that considering the huge amount of pronominal anaphors in natural language texts, the standard practice of discarding them as stop words results in loss of a significant number of distributional contexts, while resolving anaphora would provide a model with more contexts for words functioning as antecedents, optimizing data usage. This is particularly important for relatively small corpora, such as Russian National Corpus. Note that it is regularly used as a primary resource for developing computational tools for the Russian language, including distributional models.

In natural language processing, to perform anaphora resolution means to associate co-indexed elements with each other, finding their common referent. In other words, this is a task of finding the most probable antecedent for a given anaphor. It is a well-studied field of natural language processing with its own methods and tools. Unfortunately, there is evident lack of corresponding publicly available tools for Russian. For this experiment we used *An@phora*[4], an open-source tool for pronominal anaphora resolution, which has been successfully tested on the major pronominal anaphor types: relatives, reflexives, possessives and personal pronouns [Kutuzov, Ionov 2014]. It is based on a set of rules and takes as an input morphologically analyzed sentences, looking for candidate antecedents in a window of *n* words length to the left of the current anaphor. The authors claim that *An@phora* achieves precision and recall of up to 0.6 on Russian texts.

---

[2] http://ruscorpora.ru/en

[3] http://opencorpora.org

[4] http://ling.go.mail.ru/anaphora

5. Experimental setting

5.1. Resources and tools

To test our core assumption, we use full Russian National Corpus as training material. It was linguistically pre-processed with *NLTK* [Bird et al. 2009], *Mystem* [Segalovich 2003] and *An@phora* resolver. To train neural embedding models on the resulting corpora, we employed *Gensim* framework [Řehůřek and Sojka 2010], which implements CBOW and SkipGram algorithms in Python.

5.2. Anaphora resolution

Our aim is to explore the overall effect of anaphora resolution on the resulting vector models, along with type-wise comparison. To this end, antecedents for all anaphors in the training corpus were found. Then, all types of anaphors were replaced with the detected antecedents. Additionally, we produced 3 more transformed variants of each document, in which only one of 3 anaphora types recognized by our resolver was replaced. The exact list of anaphors identified by *An@phora* is given in Table 2.

Table 2. Types of Russian anaphoric pronouns detected by *An@phora*

Personal pronouns	Relatives	Reflexives
1. *он* 2. *она* 3. *оно* 4. *они* 5. *его* 6. *ее* 7. *их* 8. *мой*	1. *который*	1. *себя* 2. *свой*

The length of analysis window (the distance at which an antecedent was searched) was set to 23 words, as per recommendations of *An@phora* authors. Table 3 provides statistics of anaphora replacement performed in the Russian National Corpus.

Table 3. Total number of anaphors and resolved anaphors in RNC

Anaphor type	Total occurrences	Number of resolved anaphors (antecedent found)
Reflexives	1,200,079	1,105,853 (92.15%)
Relatives	789,080	743,018 (94.16%)
Personal pronouns	5,805,519	4,412,671 (76.00%)

As a result, the majority of anaphoric pronouns were replaced with their antecedents (mostly nouns and noun phrases). See the example below, with the original sentence as *(a)*, and the resulting one as *(b)*:

(2) a) *Речь идёт о том, что* ***коллектив***$_k$ *должен нести ответственность за результаты* ***своей***$_k$ *деятельности и выступать продавцом* ***своих***$_k$ *услуг на рынке.*
The matter is that a company$_k$ *must take responsibility for the results of* ***its***$_k$ *activity and act as a vendor of* ***its***$_k$ *services.*
b) *Речь идёт о том, что* ***коллектив*** *должен нести ответственность за результаты* ***коллектив*** *деятельности и выступать продавцом* ***коллектив*** *услуг на рынке.*
The matter is that a ***company***$_k$ *must take responsibility for the results of* ***company***$_k$ *activity and act as a vendor of* ***company***$_k$ *services.*

The example illustrates how anaphora was successfully resolved for the word *'свой'* ('its'). Note that the difference in word forms is eliminated during lemmatization. As a result, the model learned closer connection between the words *'коллектив'* ('company'), *'услуги'* ('services') and *'деятельность'* ('activity').

In case no antecedent was found for the given anaphor, the latter was discarded as a stop word. Less than 10% of all texts in the corpus (mainly, short news bulletins and jokes) contained no anaphoric relations. Note that the sentences with resolved anaphora replaced the original ones, so the raw amount of training material did not change significantly.

5.3. Preprocessing

Four training corpora were produced during the previous stage: one for each of the 3 anaphora types and one with all identified anaphors replaced. The 5th one is the baseline control corpus with no anaphors replaced. Preprocessing included sentence splitting (with *NLTK*), tokenization, lemmatization, PoS-tagging (with *Mystem*) and stop words removal. For the latter we used NLTK default stop list for Russian. Numeric tokens and punctuation were also discarded.

5.4. Models training

We trained distributional models with the following fixed hyperparameters:

- Vector size: 300;
- Minimal frequency to consider a word during training: 3;
- Negative samples: 15.

They were chosen as the best for RNC models ([Kutuzov, Andreev 2015]).

We experimented with tuning the following hyperparameters:

1. Learning algorithms: SkipGram or CBOW;
2. Width of symmetrical context window: 1, 2, 3, 5, 10, or 20 words.

As we have 5 training corpora (control `baseline' corpus, all anaphors replaced, only personal pronouns replaced, only relatives replaced and only reflexives replaced), 2*6*5=60 models were trained all in all.

6. Evaluation

6.1. Methods

The task of measuring semantic relatedness is fundamental for distributional semantics and is a good test for the results of our experiment. If word vectors trained on corpora with anaphora resolved gained consistent increase in quality, it means that the experiment was successful and anaphora resolution does increase the models' performance. The results were evaluated against two gold standards for measuring semantic similarity in word pairs: **RUSSE'15** training set [Panchenko et al. 2015] and **Simlex-999** [Hill et al. 2015].

In the first case using the training dataset instead of the test dataset allows for more sound statistics due to larger number of word pairs with annotated measure of relatedness: 209,320 versus 14,836 in the test set. RUSSE evaluation standard suggests four tasks: **hj** (Spearman's correlation with expert annotations), **rt** (average precision for word pairs from *RuThes Lite*), **ae** (average precision for associations from the *Russian Associative Thesaurus*) and **ae2** (average precision for associations from the *Sociation.org*).

6.2. Results

In Table 4 we show the difference between the best baseline models trained on raw RNC (no anaphora resolution) and those RNC versions where anaphora was resolved. We also provide training parameters for every model (learning algorithm and window size). Full tables with all the results can be found in the Appendix[5].

[5] http://ling.go.mail.ru/misc/dialogue_2016.html

Table 4. Performance in RUSSE tasks for the best experimental models

RUSSE task	Best raw models with lemmatization and PoS-tagging	Best anaphora-enriched models
hj	0.75608 CBOW / Window 10	**0.76529** Reflexives replaced CBOW / Window 20
rt	0.78311 CBOW / Window 1	**0.78348** Relatives replaced CBOW / Window 1
ae	0.83045 SkipGram / Window 20	**0.83899** All anaphors replaced SkipGram / Window 20
ae2	0.85341 CBOW / Window 20	**0.86660** All anaphors replaced CBOW / Window 20

Improvements are obvious: in all tasks, the models trained on corpora with anaphora resolution were the best. The most significant benefit was seen in **hj** and associative **ae2** tasks. Relatedness task **rt** demonstrated the least evident change. In most tasks CBOW algorithm worked best while in **ae** Skip-gram algorithm was a better option. However, anaphora resolution everywhere resulted in stable, parameter-independent quality growth. We could not single out one best combination of training parameters and anaphora type to resolve in corpora, so it seems to be task-dependent.

Additionally, we evaluated all the models which performed best in the RUSSE tasks against Simlex-999. This is the Russian section of multilingual human-annotated dataset, consisting of 999 word pairs manually annotated with semantic similarity. The results are summarized in Table 5.

Table 5. Spearman correlation against SimLex-999 gold standard

Model	Spearman correlation
All anaphors replaced CBOW / window 20	**0.613**
Raw corpus CBOW / window 20	0.606
Reflexives replaced CBOW / window 20	0.604
All anaphors replaced SG / window 20	0.599
Raw corpus CBOW / window 10	0.597
Raw corpus SG / window 10	0.589

Even this relatively sparse test data confirms the advantage which anaphora resolution lends to distributional models. It is interesting that resolution of personal pronouns alone did not result in high accuracy. In our opinion, this is a backlash of relatively low quality of our current anaphora resolver. The majority of wrong antecedents were assigned to personal pronouns. Along with their high frequency, this significantly impaired the results.

Resolution of all anaphors gave boost to two associative metrics, which normally benefit from large window parameter. Anaphora resolution generally increases the number of words in sentences, thus it is natural that large windows also allowed anaphora to take greater effect. The same explanation can be offered for relatively small improvement in case of **rt**, where the best models were those with the smallest window possible.

Table 6 and Table 7 demonstrate the performance of 4 best models trained on raw corpora with minimal preprocessing (only lower-casing tokens). This metrics illustrates the extent to which linguistic analysis as a whole improves model performance.

Table 6. RUSSE metrics of the best models trained on raw corpus with minimal preprocessing

RUSSE task	CBOW / window 1	CBOW / window 10	CBOW / window 20	SG / window 20
hj	0.26680	0.47501	0.54553	0.58660
rt	0.63063	0.71158	0.71750	0.67318
ae	0.54013	0.66039	0.67387	0.67318
ae2	0.54013	0.78991	0.80959	0.67318

Table 7. Spearman correlation against SimLex-999 gold standard (the best models trained on raw corpus with minimal preprocessing)

CBOW / window 1	0.2538
CBOW / window 10	0.3982
CBOW / window 20	0.4311
SG / window 20	0.5093

7. Conclusions and future work

In this paper we employed pronominal anaphora resolution as a way to optimize data usage for training word embedding models. In the course of our experiments, we trained neural distributional models on Russian National Corpus with resolved anaphors of the following types: personal pronouns, relatives, and reflexives. The performance of the resulting models in various semantic tasks was then evaluated. We showed that anaphora resolution results in evident improvement of models' performance. With the same input and identical set of hyperparameters, 'anaphora-enriched' models were consistently ahead their baseline competitors by several points.

These results are especially promising, considering that the employed anaphora resolution tool (*An@phora*) is not perfect: with 60% accuracy, a good part of the detected antecedents is wrong, and for some anaphors, antecedents are not found at all. Nevertheless, a consistent increase in semantic similarity task performance is observed.

Unfortunately, no other publicly available anaphora resolution tools for Russian exists yet. Thus, we were unable to estimate how much the resolver performance influences the results. We look forward to new upcoming tools, which would make such an experiment possible.

With corresponding instruments at hand, it would also be interesting to resolve not only anaphora but also coreference chains, to check if it provides additional performance boost. We assume this a more advanced problem, since mere replacement will not work: getting new contexts for some units will mean losing those for others.

Anaphora is a universal linguistic phenomenon, so our work can be applied to other languages as well. More tools for anaphora resolution are available for some languages, cf. for example, Stanford Deterministic Coreference Resolution System [Recasens et al. 2013]. Thus, experiments on English are in our nearest plans.

References

1. *Adel H., Schütze H.* (2014), Using Mined Coreference Chains as a Resource for a Semantic Task. In EMNLP (pp. 1447–1452).
2. *Baroni, M.* (2008). Distributional Semantics (slides)
3. *Baroni M., Dinu G., Kruszewski, G.* (2014), Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 238–247.
4. *Bird S. Loper E., Klein E.* (2009), Natural Language Processing with Python. O'Reilly Media Inc.
5. *Kevin Clark* (2015), Neural coreference resolution. Stanford Report
6. *Firth J. R.* (1957), "A synopsis of linguistic theory 1930–1955". Studies in Linguistic Analysis (Oxford: Philological Society): 1–32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952–1959. London: Longman.
7. *Hill F., Reichart R., Korhonen A.* (2015), Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.
8. *King Jeffrey C.* (2013), "Anaphora", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.).
9. *Kutuzov A., Ionov M.* (2014), The impact of morphology processing quality on automated anaphora resolution for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (Bekasovo, June 4–8, 2014), issue 13 (20) , Moscow, RGGU.
10. *Kutuzov A., Andreev I.* (2015), Texts in, meaning out: Neural language models in semantic similarity tasks for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (Moscow, May 27–30, 2015), issue 14 (21), Moscow, RGGU.

11. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems.
12. *Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N.* (2015), RUSSE: The First Workshop on Russian Semantic Similarity, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". (Moscow, May 27–30, 2015), issue 14 (21), Moscow, RGGU.
13. *Pekar, V.* (2004, August). Linguistic preprocessing for distributional classification of words. In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries (pp. 15–21). Association for Computational Linguistics.
14. *Plungian V. A.* (2005), Why we make Russian National Corpus? [Зачем мы делаем Национальный корпус русского языка?], Otechestvennye Zapiski, 2
15. *Řehůřek R., Sojka P.* (2010), Software framework for topic modeling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta.
16. *Segalovich I.* (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, MLMTA, pp. 273–280.
17. *Recasens, M., de Marneffe, M. C., & Potts, C.* (2013). The Life and Death of Discourse Entities: Identifying Singleton Mentions, HLT-NAACL (pp. 627–633).