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

Moscow, June 1–4, 2016

ESTIMATING SYNTAGMATIC ASSOCIATION STRENGTH USING DISTRIBUTIONAL WORD REPRESENTATIONS[1]

Bukia G. T. (gregorybookia@yandex.ru),
Protopopova E. V. (protoev@yandex.ru),
Panicheva P. V. (p.panicheva@spbu.ru),
Mitrofanova O. A. (o.mitrofanova@spbu.ru)

St. Petersburg State University, St. Petersburg, Russia

Abstract: In the paper we present distributed vector space models based on word embeddings and a specific association-oriented count-based distributional algorithm which have been applied to measuring association strength in Russian syntagmatic relations (namely, between nouns and adjectives). We discuss the compositional properties of the vectors representing nouns, adjectives and adjective-noun compositions and propose two methods of detecting the syntactic association possibility. The accuracy of the proposed measures is evaluated by means of a pseudo-disambiguation test procedure and all models show considerably high results. The errors are manually annotated, and the model errors are classified in terms of their linguistic nature and compositionality features.

Keywords: distributional semantics, compositional collocations, adjective-noun phrases, association measures, Word2Vec, pseudo-disambiguation, Russian corpora

[1] The reported study is supported by RFBR grant № 16-06-00529 "Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques".

ОЦЕНКА СТЕПЕНИ СВЯЗИ В СИНТАКСИЧЕСКИХ КОНСТРУКЦИЯХ С ИСПОЛЬЗОВАНИЕМ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ

Букия Г. Т. (gregorybookia@yandex.ru),
Протопопова Е. В. (protoev@yandex.ru),
Паничева П. В. (p.panicheva@spbu.ru),
Митрофанова О. А. (o.mitrofanova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Аннотация: В статье описан оригинальный подход к оценке связей в синтаксических конструкциях (прежде всего, в сочетаниях «прилагательное + существительное»). В экспериментах используются векторные модели, основанные на Word2Vec и на авторской мере оценки связей, учитывающих сочетания, не наблюдаемые в корпусе текстов. Исследовательские данные позволяют делать выводы о композиционности сочетаний и о синтаксических связях между частями сочетаний. Оценка параметров используемых нами моделей осуществляется в рамках так называемой псевдо-дизамбигуации. В ходе тестов обе модели показали высокие результаты. Мы провели анализ ошибочных разборов сочетаний и выявили несколько типов ошибок, в числе которых метафорические конструкции, конструкции с частично десемантизированными элементами, переходные случаи.

Ключевые слова: дистрибутивная семантика, композиционные сочетания, сочетания «существительное+прилагательное», меры ассоциации, Word2Vec, псевдо-дизамбигуация, русскоязычные корпусы

1. Introduction

Semantic compatibility (term used, for example, in (Ghomeshi, Massam 1994)) or the ability of two words or constructions to collocate has been widely studied within different theoretical frameworks (Goldberg 1995), (Apresjan 1974). The meaning of complex linguistic units such as collocations is generally assumed to be non-compositional so it is not fully derived from the meaning of their parts. Consider Apresjan's example of Russian adjectives 'goryachij' and 'zharkij' (Apresjan 2010), both translated in English as 'hot'. They are treated as synonyms, although in fact they are not fully interchangeable in contexts as they have virtually non-overlapping sets of collocates. 'Goryachij' is used to refer to a local sensation ('*goryachij shokolad*'—hot chocolate) and 'zharkiy' expresses the idea of a more general sensation of the environment ('*zharkoe leto*'—hot summer).

Such restrictions can be treated within Construction Grammar (Goldberg 1995) theory claiming that lexical constructions reveal the unity of form and meaning. The form is maintained by fixed elements of constructions on the one hand, and on the other hand, by selectional (morphosyntactic, lexical-semantic, propositional, etc.) restrictions imposed on the slot fillers. Construction Grammar observes a wide range of variations, from free co-occurrence of lexical features to highly idiomatic units. In our study constructions are treated as multilevel structures combining lexical (lemmata, wordforms), grammatical and semantic features. Such an approach allows us to describe collocability of a target word in a given sense in terms of constructions.

The degree of association in lexical constructions is an important factor in such NLP applications as paraphrase generation for machine translation, language modelling, automated and semi-automatic dictionary acquisition, semantic role labelling, word sense disambiguation, etc. A number of collocation extraction methods rely on corpus evidence assuming that if a construction occurs in texts, its components can be combined. However, these methods are not applicable when a pair of words is not observed in texts. Moreover, we can imagine occasional word combinations that are not generally expected in natural context (*'дремучее равновесие'*—primeval balance).

In our study we compare two approaches to measuring association strength in word combinations revealing possibility of syntagmatic relations, the first one assuming compositionality of their meaning, the second one implying that such word combinations have a meaningful unit which is not derived from word meanings. The paper is organized as follows: first of all, an outline of the research in the field is presented. Then, we describe two approaches to measuring association strength using distributed vector representations. Finally, the performance of the models is evaluated within a pseudo-disambiguation benchmark, and in conclusion a brief error analysis is presented.

2. Related work

Recently distributional semantic modelling has been applied to studying meaning of complex linguistic units (constructions, clauses, sentences) with the help of vector space models and their modifications (Kolb 2008), (Pekar, Staab 2003), (Sahlgren 2006), (Schütze 1992), (Widdows, Cohen 2010), etc. SemEval 2014 competition[2] included evaluation of compositional distributional semantic models of full sentences for English. One of the recent examples is COMPOSES which uses compositional operations to model linguistic units in semantic space. "Content" words (e.g. nouns) are represented as vectors, while relational words (e.g., adjectives) correspond to functions mapping input items to compositional structures (Baroni et al. 2014).

A survey on mathematical operations applied to determine compositional meaning is presented in (Kartsaklis, Sadrzadeh 2013). The authors focus their attention on tensor-based models where relational words (verbs, adjectives) are regarded as tensors. The distributed vector representations (Mikolov et al. 2013a) are also studied with respect to their compositionality (Mikolov et al. 2013b).

[2] http://alt.qcri.org/semeval2014/task1/

Distributional models for Russian have been applied in a number of applications. Serelex semantic model is incorporated in an information retrieval system which gets a target word as an input and gives a list of its associates as an output (Panchenko 2013). It provides contextual correlates for a target word which are ranked according to an original similarity measure based on lexical-syntactic patterns.

The evaluation of various association measures and Russian distributional models has been discussed in RUSSE competition (Panchenko et al. 2015). However, semantic relatedness evaluation involves only paradigmatic relations between lexical units. Thus, to our knowledge, there has been no evaluation of vector space models applied to syntagmatic relations in Russian.

A recent study concerning association strength measurement in syntactic constructions and testing methodology is described in (Bukia et al. 2015). The experiments are conducted on syntactic constructions. The authors train association measure on adjective-noun collocations from a very small corpus of 350 thousand sentences. The algorithm yields high performance in predicting association possibility although it is based on a small amount of training data. Their approach is detailed below and compared with association strength evaluation results produced by vector space modelling.

Association strength measurement is closely related to identification of abnormal lexical compositions (Vecchi et al. 2011) and automatic lexical error detection (Kochmar, Briscoe 2013). The latter work presents a number of semantic anomaly measures in a vector space. We adopt one of the measures and apply it to a semantic space with reduced dimensionality produced by Word2Vec.

3. Distributed word representations and their application to association measurement

3.1. Distributed word representations in word2vec toolkit

Continuous word representations in vector space have been gaining extreme popularity since (Mikolov et al. 2013a). As reported in the paper, high quality word vectors are obtained by training recurrent neural network with two different architectures—continuous-bag-of-words (CBOW) and skip-gram. Both yield considerable results in similarity and association measurement when using the cosine similarity measure. The authors implement their approach in a widely used word2vec[3] toolkit.

3.2. Distributed word vectors and linguistic regularities

(Mikolov 2013a) have proposed a questionnaire method (later elaborated in (Mikolov et al. 2013c)) to estimate word vector representations in terms of semantic and syntactic relationships between words that are learned automatically. The

[3] https://code.google.com/archive/p/word2vec/

question is formed of two pairs of words with the same relationship such as "What is the word (x) that is similar to $small(x_c)$ in the same sense as $biggest(x_b)$ is similar to $big(x_a)$?" It turns out that the vector

$$y = x_b - x_a + x_c$$

is most similar to x in terms of cosine similarity. This was proved on several groups of questions including semantic relationship (*the capital of*, *the currency of*, *the female of*, etc.) and syntactic or, more precise, grammatical ones (*past form of*, *superlative form of*, etc.).

The difference of two word vectors characterizes their relation which is independent of their own meanings and can be used to infer the missing word in a different pair representing the same relation.

This observation may be applied to measuring association in syntactic constructions even for word-pairs not attested in the corpus. We assume that if a pair of words comprises a collocation or construction, there is a regular semantic relationship between the two words, which is systematically replicated in other word-pairs attested in the corpus. Thus we can find such a pair of words in the corpus that its difference vector is similar to the corresponding difference vector of the given words. Otherwise, if the combination is unacceptable, the difference vector is unpredictable and does not have similar vectors in corpus. This difference vector often accounts for a relation which can not be formulated clearly but appears regularly in syntactic word combinations. Consider several examples of a noun + adjective combination and the nearest pair:

- 'овощной салат' (vegetable salad)—'конфетная коробка' (a box of sweets), 'гороховый суп' (pea soup);
- 'чёрный кофе' (black coffee)—'тёмное пиво' (dark beer), 'розовое мартини' (pink martini).

  The example sets consist of definitions by content and color respectively.

  The association measure for a combination (a,n) is formulated as:

$$W2V_{rel} = \max_{a_i,n_j \in K} \frac{< n, n_i - a_i + a >}{|n||n_i - a_i + a|},$$

where the maximum value is found over all pairs (a_i, n_j) occurring in the corpus. This measure is referred to as $W2V_{rel}$ below.

3.3. Compositional approach to association measurement

We adopt the compositional approach investigated in (Kochmar, Briscoe 2013), (Vecchi et al. 2011). The assumption is that the vector representing a noun-adjective composition is meaningful if it is closely related to the head of the composition, i.e. the initial noun. The similarity measure between the composition and the head noun is expected to positively reflect the acceptability of the noun-adjective association. The acceptable compositions are expected to be ranked as more similar to the initial head nouns than the unacceptable ones. However, with normalized vectors, as in Word2Vec approach, the monotonicity of the functions *Similarity1* and *Similarity2* is the same, although *Similarity2* measures the simple cosine similarity between noun and adjective:

$$Similarity1(noun, adj) = cos(noun + adj, adj)$$
$$Similarity2(noun, adj) = cos(noun, adj)$$

Quantifying similarity between the noun and the adjective in the same vector space yields here the same result as when quantifying similarity between the initial noun and the attributive noun phrase. The latter formula is linguistically motivated and naturally interpreted, which is not so obvious about the former. These values will be referred to as *Comp* below.

3.4. Count-based distributional approach

We compare the described methods to an approach proposed in (Bukia et al. 2015). Their association measure is based on distributional word properties concerning only a fixed construction, namely, the noun-adjective association (referred to as *D* below).

The basic assumption is that if two words relevant for a construction slot collocate in texts with similar words (contexts) relevant for another slot, the probability of the first target word to be combined with the contexts of the second target word and, vice versa, is high, even when some pairs are not observed in texts. This idea is formally expressed by the notion of confusion probability, which is computed as follows: given the contexts of the first word $c(x_1)$ and the second one $c(x_2)$, their confusion probability is equal to P:

$$\mathbb{P}\{x_1 \sim x_2\} = \frac{|c(x_1) \cap c(x_2)|^2}{|c(x_1)|\,|c(x_2)|}$$

The association strength between two words in a collocation occurring in a corpus is usually computed by means of Fisher's exact test (Stefanowitsch 2003). The distributional association measure between a noun and an adjective in a collocation *D*(*a*,*n*) is then defined as an average of such counts over all confusable words weighted by the confusion probability. As discussed in (Bukia et al. 2015), the highest results are produced by such a measure based on mutual information (MI) counts.

4. Evaluation

4.1. Data and experimental setup

We use a corpus of Russian fiction (146M tokens obtained from M. Moshkov's digital library, URL: lib.ru). All preprocessing (tokenization, lemmatization, shallow morphological analysis) was performed by means of PyMorphy2 Python library (URL: http://pymorphy2.readthedocs.org/en/latest/). About 157K (80K unique) adjective-noun pairs were extracted from these texts.

The 150-dimensional vectors were trained using Gensim library (Rehurek, Sojka 2010) with skip-gram architecture and 4-word window. The count-based distributional association takes into account only corpus frequencies of a noun, an adjective and their

combination. The evaluation procedure follows pseudo-disambiguation test as described in (Bukia et al. 2015). It has been also used in (Pekar 2004), (Tian et al. 2013).

The following lemmata were extracted from the corpus:

- 500 random nouns $N = \{n_i\}$;
- for each noun a random adjective a_i collocating with this noun;
- for each a the nearest adjective by frequency $X = \{x_i\}$ (not attested in combination with the corresponding noun n_i).

All combinations (a_i,n_j) are then removed and the system is trained on the rest of the corpus. Thus, 500 triplets consisting of a target noun, an acceptable and an unacceptable combination are obtained. The task is to find out, which combination of an adjective and a noun was removed, i.e. which one is acceptable but deleted from the final training corpus. In our case, the first association value is expected to be higher than the second one.

4.2. Results

It should be noticed that the pseudo-disambiguation task is limited by the fact that we do not know anything about the second combination not attested in the corpus. Thus, the results were manually checked in order to eliminate malformed triplets. In some cases, either both combinations are acceptable or even none of the chosen ones.

The accuracy counts are presented in Table 1 in the following order:

- raw result based on the assumption that the first possible pair is acceptable while the second one is not (*Acc*);
- pseudo-disambiguation accuracy calculated after manual annotation of triplets (*Corr*).

As mentioned above, the models are denoted as follows:

- $W2V_{rel}$ for vector difference based measure (Section 3.2);
- *Comp* for measure based on vector composition (Section 3.3);
- *D* for a simple distributional measure (Section 3.4).

First of all, it should be noticed that the models based on word embeddings achieve higher accuracy than a count-based one. However, the latter one has an important advantage: its results are easy to implement and interpret.

Secondly, the results presented below should be compared only with the data provided by the models performing the same task: estimating association strength for unseen combinations. The most recent work (Tian et al. 2013) based on quite different principles reports 88% accuracy. Thus, the discussed models yield state-of-the art performance.

Table 1. Accuracy and real error percentage in the pseudo-disambiguation task

	$W2V_{rel}$	*Comp*	*D*
Acc	76%	81%	75%
Corr	88%	93%	84%

4.3. Error analysis

After manual error annotation the errors of different models are compared. Common errors, i.e. shared by all three methods, can be divided into two groups concerning their source: acceptable combinations representing rare or occasional metaphorical expressions (*'информационная чума'—informational boom*, *'круглая сирота'—a total* (literally *'round'*) *orphan*, etc.) and those containing a word with a vague or general meaning (*'европейский квартал'—european quarter, 'серьезное ухаживание'—earnest courting*, etc.).

The rest of the errors, i.e. the model-specific ones, were ordered by the acceptable combination frequency. In each experiment a group of very rare (acceptable) combinations can be found: *'суеверный закон'—superstitious law, 'безработный фанатик'—unemployed fanatic*. These expressions are either rare themselves or contain a rare word, and even a native speaker is scarcely able to construct a sentence where these combinations are justified. The top combinations are constructions in the sense of Construction Grammar (Goldberg 1995): their meaning is not additive and can not be simply derived from the meaning of the constituents: *'жевательная резинка'—chewing gum*, *'трезвая голова'—reasonable person* (literally *'sober head'*).

The middle of these ordered lists contains real errors which are due to an algorithm structure or its assumptions: *'копировальный центр'—copy center*, *'сумасшедшая история'—mad story*. These combinations are less idiomatic and are constructed regularly. The Word2Vec compositional similarity measure (section 3.3) fails to extract such combinations because the constituents appear to have too few intersecting contexts. The errors of the count-based distributional model may also be explained by the underlying assumption that similar words occur in similar noun-adjective contexts. On the other hand, such expressions are correctly processed by the Word2Vec relative measure (see section 3.2), meaning that analogous regular relations between attested nouns and adjectives were observed when looking for the nearest difference vector. Several examples are presented in table 2.

Table 2. Examples of nearest difference vector combinations

test combination	**nearest difference combination**
жевательная резинка—chewing gum	*кавказский хребет—Caucasian chain* *цементная ступенька—cement step*
копировальный центр—copy center	*патрульный корабль—patrol ship*

Finally, it should be mentioned that although Word2Vec compositional measure has shown the best results, it can not be improved, because its assumption is not applicable in all real world cases. The accuracy of Word2Vec relative measure, on the contrary, can be increased, since the core idea of an additional meaning can be observed in real data.

5. Conclusion

We have applied the task of measuring association strength between Russian nouns and adjectives to compare compositional and relative Word2Vec semantic models with a simple distributional association measure. The test was conducted following a conventional pseudo-disambiguation methodology. The models were trained with a 11M sentences corpus where all in-sentence co-occurrences of the word pairs are deleted.

Both measures based on Word2Vec models outperformed a simpler count-based one and achieved state-of-the-art accuracy. The error analysis allows us to talk about future improvements by applying a more sophisticated measure to determine a syntagmatic relation. More exactly, we are going to focus on the interpretation of the difference vector. Another important concern is the testing methodology which is also subject to future investigation and improvement based on human judgements. For example, separate datasets for compositional and idiomatic combinations should be created and manually assessed.

References

1. *Apresjan, Ju. D.* (1974), Leksicheskaja semantika. [Lexical semantics], Moscow, Nauka.
2. *Apresjan, Ju. D.* (ed.). (2010), Prospekt aktivnogo slovarya russkogo jazyka [The prospect of active Russian dictionary)], Moscow.
3. *Baroni, M., Bernardi, R., Zamparelli, R.* (2014), Frege in space: A program of compositional distributional semantics, Linguistic Issues in Language Technology, Vol. 9, CSLI Publications.
4. *Biemann, C.* (2007), Unsupervised and knowledge-free natural language processing in the structure discovery paradigm, PhD thesis, University of Leipzig.
5. *Bukia, G., Protopopova, E., Mitrofanova, O.* (2015), A corpus-driven estimation of association strength in lexical constructions, Sergey Balandin, T. T., Trifonova, U.(eds.), Proceedings of the AINL-ISMW FRUCT, FRUCT Oy, Finland, pp. 147–152, http://fruct.org/publications/ainl-abstract/files/Buk.pdf
6. *Ghomeshi, J., Massam D.* (1994), Lexical/syntactic relations without projection, Linguistic Analysis, Vol. 24, Issues 3–4, pp. 175–217.
7. *Goldberg, A.* (1995), Constructions: A construction grammar approach to argument structure, University of Chicago Press, Chicago.
8. *Kartsaklis, D., Sadrzadeh, M., et al.* (2013), Prior disambiguation of word tensors for constructing sentence vectors, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNL), pages 1590–1601, Seattle, USA.
9. *Kochmar, E., Briscoe, T.* (2013), Capturing anomalies in the choice of content words in compositional distributional semantic space, Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, pp. 365–372.
10. *Kolb, P.* (2008), Disco: A multilingual database of distributionally similar words, Proceedings of KONVENS-2008, Berlin.

11. *Mikolov T., Chen K., Corrado G., Dean J.* (2013a), Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at International Conference on Learning Representations.
12. *Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.* (2013b), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, pp. 3111–3119.
13. *Mikolov, T., Wen-tau Yih, Zweig, G.* (2013c), Linguistic Regularities in Continuous Space Word Representations, Proceedings of NAACL HLT, pp.746–751.
14. *Panchenko, A., Loukachevitch, N., Ustalov, D., Paperno, D., Meyer, C., Konstantinova, N.* (2015), Russe: The first workshop on Russian semantic similarity, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Moscow.
15. *Panchenko, A., Romanov, P., Morozova, O., Naets, H., Philippovich, A., Romanov, A., Fairon, C.* (2013), Serelex: Search and visualization of semantically related words, Advances in Information Retrieval, Springer Verlag, pp. 837–840.
16. *Pekar, V., Staab, S.* (2003), Word classification based on combined measures of distributional and semantic similarity, Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pp. 147–150.
17. *Pekar, V.* (2004), Distributivnaja model sochetaemostnyh ogranichenij glagolov [A distributional model of verbal selectional restrictions]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2004" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2004"], Moscow.
18. *Rehurek, R., Sojka, P.* (2010), Software framework for topic modelling with large corpora, Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, Valletta, Malta, pp. 46–50, 2010.
19. *Sahlgren, M.* (2006), The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, PhD thesis, University of Stockholm.
20. *Schutze, H.* (1992), Dimensions of meaning, Proceedings of Supercomputing'92, pp. 787–796.
21. *Stefanowitsch, A., Gries, S. T.* (2005), Collostructions: Investigating the interaction of words and constructions, International journal of corpus linguistics, Vol. 8(2), pp. 209–243.
22. *Tian, Z., Xiang, H., Liu, Z., Zheng, Q.* (2013), A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation, Proceedings of the ACL Meeting, 2013, pp. 1169–1179.
23. *Vecchi, E. M., Baroni, M., Zamparelli, R.* (2011), (Linear) maps of the impossible: capturing semantic anomalies in distributional space, Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 1–9.
24. *Widdows, D., Cohen, T.* (2010), The semantic vectors package: New algorithms and public tools for distributional semantics, IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 9–15.