ON THE NATURE OF SEMANTIC SIMILARITY AND IT'S MEASURING WITH DISTRIBUTIONAL SEMANTICS MODELS

Ryzhova D. A. (daria.ryzhova@mail.ru),
Kyuseva M. V. (mkyuseva@gmail.com)

NRU Higher School of Economics, Moscow

The paper describes our application of the distributional semantic model (DSM) method that we developed for The First International Workshop on Russian Semantic Similarity Evaluation (RUSSE) shared relatedness task. The model was trained, for the most part, on the data of the Russian National Corpus main subcorpus (around 200 mln tokens), and the resulting vector space was weighted according to "ppmi" (Positive Point-wise Mutual Information) and "plmi" (Positive Local Mutual Information) weighting schemes.

The results of the workshop show that classical distributional semantic models trained on relatively small corpora can provide data of high quality.

Keywords: distributional semantic models, composition models, semantic similarity

1. Introduction

This paper results from our participation in The First International Workshop on Russian Semantic Similarity Evaluation (RUSSE, http://russe.nlpub.ru/) shared task. The system we present has been developed for the relatedness task, whose goal is to capture synonyms, hyperonyms ans hyponyms.

In Section 2 the main points of the system organization are described: subsection 2.1 gives general information, subsection 2.2 discusses the corpora used, subsection 2.3 tells about the weighting schemes and subsection 2.4 discusses some challenges we met. In Section 3 a brief conclusion is presented.

2. System structure

2.1. General information

Our system is based on the distributional semantics model (DSM) techniques (Baroni et al. 2013). DSMs come from the hypothesis that the meaning of a lexical item can be expressed via contexts it is used in. Lexical meaning in these models is represented as a multidimensional vector whose components are a function of the number of times the word occurred in a certain environment in the corpus. This framework has already proved to be useful in accomplishing similar tasks, i.e. in forming lists of synonyms and co-hyponyms (Landauer & Dumais 1997).

Our system consists of vector representations for the words from the task dataset. As vector components we used 10,000 most frequent Russian nouns, verbs, adjectives and adverbs. Frequencies were computed on the basis of the main subcorpus of the Russian National Corpus that we used to train the model (see below).

The value of every component is the absolute number of occurrences of a given component word within a window ±5 (5 content words to the left and 5 content words to the right) around the target lexeme. Sentence boundaries are not taken into account.

2.2. Text corpora

The main subcorpus of the Russian National Corpus (~220 mln tokens, www.ruscorpora.ru) was used for collecting the most part of vectors. Presented in the form of a plain text, it should be preprocessed in order to operate with lexemes instead of word forms. We provided their morphological annotation with the help of Mystem system (Segalovich 2003) and then used a morphological disambiguator constructed by MA students of Higher School of Economics within the "Engineering of linguistic resources and systems" course (Lakomkin et al. 2013). The relatively small volume of the corpus is due to two factors. First, such an amount of data can be quickly processed by the machine. Second, as our previous research shows, this RNC subcorpus suffices to provide quality distributional semantic representations (see Kyuseva 2014, Ryzhova 2014, Ryzhova et al. to appear)[1].

The goal of the previous project was to determine whether the main lexical regularities stated on the typological data reveal themselves on the material of one language. We focused our study on frequent adjectival domains denoting physical qualities: 'sharp', 'smooth' and 'straight'. After collecting vectors for Russian noun phrases including these adjectives (*ostryj mech* 'sharp sword', *ostraya bolj* 'acute pain', *pryamaya linia* 'straight line' and so on), we computed cosine similarity metrics for all NP pairs in the sample. Comparison of these similarities with the typological data, presented in the format of the Typological database of qualitative features (Kyuseva et al. 2013) demonstrates that the closer the vectors are, the more likely the corresponding noun phrases would be translated into some other language with the help of one adjective for both phrases. For example, the Russian phrases *ostryj nozh* 'sharp knife' and *ostryj mech* 'sharp sword' (cos = 0.96) are more likely to be translated with the same adjective than *ostryj nozh* 'sharp knife' and *ostraja problema*' 'vexed problem' (cos = 0.4).

In the experiments three text corpora were used in different combinations:

1) RUWAC (http://corpus.leeds.ac.uk/internet.html)—1 billion tokens;
2) Main subcorpus of the Russian National Corpus—220 million tokens;
3) Media subcorpus of the Russian National Corpus—200 million tokens.

[1] See also similar ideas in Kutuzov&Kuzmenko to appear.

These corpora contain texts of different genres. While the first and the third corpora are collections of internet and media texts respectively, the second one represents a well balanced sample of texts in Russian.

The table below illustrates the results of the comparison between vector and typological data for two different semantic fields: 'sharp' and 'smooth'. All DSM parameters are the same, except the text corpora used.

Table 1. The main RNC subcorpus suffices to provide quality distributional semantic representations

		Results of the comparison (Pearson correlation)	
Corpus	**Corpus size**	**'sharp'**	**'smooth'**
RNC, main	~220 mln tokens	0.764	0.905
RNC (main+media) + RUWAC	~1,420 mln tokens	0.764	0.833

As it can be seen from the table, increase of the corpus volume does not significantly change the result of the comparison. In case of the 'smooth' dataset, it even decreases the correlation coefficient, which is probably due to the genre bias of the RUWAC and the RNC media corpora.

Although the main RNC subcorpus proved to be sufficient for computing quality vector representations, it has few or no examples of new colloquial and slang expressions, which forced us to use a different source of these data. These expressions, making up for about 1.6%[2] of the total sample (see some examples below), were processed on the basis of the RUWAC corpus, which contains more modern texts.

Table 2. Examples of absent words in the RNC main subcorpus

Word	**Reason of absence / low frequency**
nik 'nickname', haking 'hacking'	borrowed words used primarily in internet communication
adresatka 'female addressee'	a highly colloquial and occasional form that is never mentioned in dictionaries
podbassejn 'subbasin', elektrogazosvarsh'itsa 'female electro-gas welding operator'	rare lexemes
Avatar	proper name, film title
Sanepidblagopoluchije	proper name, name of an organization

[2] This percentage does not take into account multiword expressions, word forms in oblique cases and hyphenated words also lacking in the main subcorpus of the RNC. These special cases are treated in section 2.4.

The total number of the items that are not presented in the RNC main subcorpus makes up for about 7.7% of the test sample.

2.3. Weighting schemes

After having computed vector representations, we created a semantic space. Then, on the basis of the word frequencies of this space we performed several weighting transformations (with the help of the DISSECT toolkit, see Dinu et al. 2013). The toolkit suggests four weighting schemes: "ppmi" (Positive Point-wise Mutual Information), "plmi" (Positive Local Mutual Information), "epmi" (Exponential Point-wise Mutual Information) and "plog" (Positive Log Weighting). Using the training dataset of human judgements provided by the organizers of the workshop as the golden standard, we determined the best schemes for our purposes.

Table 3. Comparison of the weighting schemes

Weighting scheme	Spearman correlation	Pearson correlation
plog	0.07	0.06
epmi	0.595	0.546
ppmi	0.615	**0.646**
plmi	**0.719**	0.563

As the table shows, the "plmi" and "ppmi" schemes give the best results. So, we provided two variants of the system: one using the "plmi" weighting scheme, and the other—the "ppmi" transformation.

Finally, we used the cosine similarity between the corresponding vector representations as the metrics of semantic relatedness between the items from the RUSSE test sample.

2.4. Challenges

Despite the overall effectiveness of the system, a number of problems occurred.

1) Some of the lexemes from the sample (for example, *akhvashka* 'Akhvakh woman', *kinovideoteatr* 'film-videotheatre') did not appear in our text collections at all. These words make up for 0.3% from the whole dataset (29 items).

   When computing the cosine similarity metrics we kept the vectors of these words unchanged, that is, with all components set to zero.
2) The version of Mystem system that we used divided all hyphenated words into separate lexemes (for example, the word *jugo-vostok* 'south-east' was divided into *jug* 'south' and *vostok* 'east') that prevented us from computing a vector representation for such items. In these cases, we collected the statistics based on the RUWAC corpus, because it had already been annotated and disambiguated and presented such forms as indivisible lexemes.
3) Some of the hyphenated words from the test dataset presented occasional compounds that were not registered in dictionaries and were likely to be divided into different lexemes (for example, *muzykant-instrumentalist* 'musician-instrumentalist'). For these words we computed different vector representations

(a separate representation for each part of the word form) and then calculated an arithmetic mean for the values of all dimensions, thus constructing an averaged vector representation for the whole item.

4) The last two difficulties stem from the very intention of the workshop: the same test sample was designed for two different tasks (word relatedness and association between items). While for the first task it is natural to take a single word in the main form, the second one does not presuppose these restrictions. As the result, the test sample contains, besides the lexemes in the "citation form":
   (a) word forms in oblique cases (consider, for example, the pair *ogon'* 'fire' and *ljubvi* 'of love');
   (b) multiword expressions (like *pogoda* 'weather' and *na zavtra* 'for tomorrow').

   In case (a), we solved the problem with the help of morphological parser (Mystem): the lexeme was taken from the morphological analysis. In case (b), we just put a zero value of semantic relatedness for all pairs containing a multiword item.

3. Conclusion

The system described is relatively simple and builds on the classical distributional semantic models technique. As the results of the RUSSE shared task show, it is ranked high among the rivals, standing the competition with more elaborated models based on the neural networks. This brings hope that traditional DSMs are still in play and should not be untimely abandoned.

References

1. *Agirre, E. and P. G. Edmonds* (2006), Word Sense Disambiguation: Algorithms and Applications. Springer.
2. *Baroni, M.; R. Bernardi; R. Zamparelli* (2013), Frege in Space: A Program for Compositional Distributional Semantics, Linguistic Issues in Language Technologies, Vol. 9. CSLI Publications.
3. *Dinu, G., N. T. Pham, M. Baroni* (2013), DISSECT: DIStributional SEmantics Composition Toolkit, Proceedings of ACL (System Demonstrations), pages 31–36, Sofia, Bulgaria.
4. *Kutuzov A. and Kuzmenko E.* Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: the Case for Russian. To appear.
5. *Kyuseva M. V.* (2014). Verification of the Frame Approach to Lexical Typology with Distributional Semantic Models [Verifikatsiya frejmovogo podkhoda k leksicheskoy tipologii s pomoshch'yu modeley distributivnoy semantiki]. MA thesis. Higher School of Economics.
6. *Kyuseva, M. V., Reznikova, T. I., Ryzhova, D. A.* (2013), A typologically oriented database of qualitative features [Tipologicheskaya baza dannykh ad"ektivnoy leksiki], Computational Linguistics ans Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika

I Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], volume 1, Moscow, pages 419–430.

7. *Lakomkin E. D., Puzyrevskiy I. V., Ryzhova D. A.* (2013), Analysis of statistical algorithms of morphological disambiguation for texts in Russian [Analiz statisticheskikh algoritmov snyatiya morfologicheskoy omonimii v russkom yazyke] // In Proceedings of the Russian science conference AIST'2013 [Doklady vserossiyskoy nauchnoy konferentsii AIST'2013], Yekaterinburg, pp. 184–196.
8. *Landauer, T.; Dumais, S.* (1997), A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104(2), pp. 211–240
9. *Lin, D. and P. Pantel.* (2001), DIRT—Discovery of Inference Rules from Text. Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323–328
10. *Ryzhova, D. A.* (2014), Constructing of lexical typological questionnaire with distributional semantic models [Postroenie leksiko-tipologicheskoy ankety s pomoshch'yu modeley distributivnoy semantiki]. MA thesis. Higher School of Economics.
11. *Ryzhova D., Kyuseva M., and Paperno D.* Typology of Adjectives as a Benchmark for Compositional Distributional Models. To appear.
12. *Schütze, H.* (1998), Automatic word sense discrimination. Computational linguistics 24(1), pp. 97–123.