Evaluation of Twitter as semantic analysis resource

Abstract

In this paper we evaluate the use of Russian-speaking Twitter segment as a resource for distributional semantic analysis. Though gold-standard for task of semantic relatedness of Russian language does not yet exist, some tweaks may be done to substitute it. There are several gold-standards for English, German and other languages. We translate WordSim353 and evaluate it against two corpora - the one of Twitter stream and corpus made of contemporary Russian literature.

1 Introduction

One of the vital goals in task of distributional semantics of Russian language is building a gold-standard, analogous to WordSim353 (Finkelstein et al., 2001) and others. Such datasets are composed with use of expert knowledge. At first, researcher composes pairs of words, either related somehow (meronymy, hyponymy, hyperonymy, antonymy, etc.), or unrelated. Next, a group of individual experts are given the task to estimate relatedness of those pairs. Estimations may vary, e.g. four distinct grades, or fraction from 0.0 to 1.0. Typical group size is about 10-15 people.

Apperently, researchers wanting to make such datasets for other languages may try to use existing ones as basis, e.g. by simply translating them, because reengineering it from scratch would be much more complicated. Some did try, but as far as we know there was no proper evaluation of translated word-sim datasets.

Information Retrieval in a whole has been successfully applied to many languages (by Google, Yandex[1] and others). Semantic analysis was under study for decades, starting with Latent Semantic Analysis (Landauer et al., 1998), Latent Semantic Indexing, finally Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007), and numerous Wordnet and Wikipedia-based works (Zesch et al., 2008).

These may be broadly divided into two groups: ones based on graph models and measurements (e.g. Wikipedia-based), and distributional ones (LSA, LSI).

Early models, like LSA, suffered from increasing number of documents and their terms. The core feature of LSA, singular value decomposition, stated as the one to reduce vector-space model of term-document matrix, actually makes it unfeasiable because of expensive SVD operation. Tremendous shift in field of vector-space models for natural language processing was made by (Mikolov et al., 2013), with use of simplified neural network models, Continious Bag-of-Words and Continious Skip-gram. It was evaluated on word-similarity task.

On the other hand, increasing popularity of social networks, and in particular Twitter, enables users to communicate instantly, and researchers to analyse their activity. The area is "hot ground", there are works such as elections prediction (Metaxas et al., 2011), senti-

[1] www.yandex.com

Alias	Word count	WordSim, ρ_P
01	1.2M	0.25
01_10	12.5M	0.27
01_20	22.5M	0.28
books	450K	0.31

Table 1: Corpus under study

ment analysis and so on.

1.1 Goals of this paper

In this paper we focus on word-to-word similarity task for Russian language. The primary goal is to evaluate applicability of Twitter stream in this NLP task. The secondary goals are construction and validation of word similarity dataset for Russian based on WordSim353 translated subset, and cross-checking it with corpus of Russian literature.

2 Input data and algorythm

2.1 Input data

Twitter data is mined from Twitter streaming API. It is assumed to be random subset of actual Twitter stream.

Big picture of algorythm:

- fetch tweets
- filter non-Cyrillic ones
- stemming (each tweet as one sentence)
- remove stopwords
- store words in database.

All fetched data is stored in daily chunks, appox. 450k Tweets per day.

We detect Russian words by simply counting cyrillic symbols. We use Yandex Tomita Parser[2] for stemming.

Algorythm for parsing books slightly differs: it splits continious text by sentence.

2.2 Counting distributions

The word distribution matrix has word-by-word structure. Each row is frequence distribution of word context. Every cell X in row Y describes number of sentences with both x and y.

Each cell in matrix is then weighted with entropy: $-\sum p \log p$, as stated by (Landauer et al., 1998).

Semantic relatedness is computed as cosine similarity between word distributions (which is common for such task). We will ise notation where 0.0 is no similarity, and 10.0 is identically similar, as in WordSim353.

3 Evaluation

Common approach for evaluation of word semantic relatedness is to use several datasets (3-4 typically). Since we want to evaluate if translated one is worthwhile, it's enough to translate just one. We make an asdumtion that bias of translation would be more significant than the one of a dataset.

3.1 Method of translation

First of all, we consider only 2000 most frequent words in our Twitter corpus. We make list of words from WordSim353 combined set. For each word we manually lookup translation with dictionary [3] and write down short-list of possible translations. If translation is not present in 2000 list, we remove it. Empty translation lists are removed. Pairs from WordSim with partial or none translation are also removed.

Still we managed to translate roughly 250 words out of 450, giving us 100 translated pairs. This is the baseline for all our subsequent experiments.

4 Experimental results

Twitter corpus under study consists of 20 chunks since 1 to 20 August, 2014, one chunk per day. Each chunk contains $\approx 1.3M$ words. We also consider joined chunks, namely 01_10, 11_20 containing 10 days each, and 01_20 containing entire corpus. We use Pearson correlation coefficient to estimate accuracy of our method.

[2] https://tech.yandex.ru/tomita/

[3] slovari.yandex.ru

#	Pair		$m_{WordSim353} - m_{\mathtt{books}}$	$m_{WordSim353} - m_{\mathtt{01}_{2}\mathtt{0}}$
1	psychology (психолог)	depression (депрессия)	6.25303	-0.14052
2	precedent (случай)	group (группа)	-3.1325	-7.11912
3	network (сеть)	hardware (техника)	7.21652	0.35387

Table 2: Estimation errors

First measurements of Pearson correlation for single-day chunk, witout using stopwords filtering, showed as low as 0.20, in contrast with 0.6 being state-of-the-art accuracy for this task (Mikolov et al., 2013).

This low accuracy may be in result of following factors:

- errors in translation
- algorythm issues
- corpus quality
- size of corpus

Usage of stopwords helped us to make 0.25.

Next obvious step was to determine if we can improve accuracy with just more data.

It gave us another 2-3% (Table 1), with 20 times larger corpus. Apparently, enlarging it further doesnt make much sense.

Still we had to mitigate possible translation issues and corpus-related ones.

4.1 Russian literature corpus

To overcome translation issues we mined different corpus, using publicly available words of Russian contemporary authors (late 20th century - beginning of 21th, for list of them see Appendix A). In this paper we address it `books`.

Size of corpus was chosen empirically: we stopped adding texts when no significant improvement in accuracy was be noticed.

It's top was around 0.30, which is slightly better than `01_20`, and much better than `01`. Intercorrelation between `01_20` and `books` appeared to be 0.63 (on the set of pairs from translated WordSim).

From this point we may make some conclusions about quality of Twitter as lingustic resource. First of all, pretty strong correlation with `books` corpus lets us state that they have much in common, and Twitter may be used as linguistic resource.

On the other side, considering that `books` is 50 times smaller than `01_20`, it's closest contestant with 0.28 accuracy, one may conclude that literature is more information-dense linguistic resource. Which is not surprizing.

4.2 Language issues

In order to validate our word translations, we count errors for estimated relatedness values. Some of them can be seen at Table 2, Russian translation comes in parenthesis.

"Psychology" was translated as "psychologist", the *occupation*, not the *science*, "hardware" — as general "technics", and "precedent" as "case", rather than, literally, precedent.

After adjusting translation for these three words, we measured accuracy again, and saw significant shift. Firstly, $\rho_{\mathtt{books}}$ jumped to 0.39, but somehow $\rho_{\mathtt{01\_20}}$ bumped down to 0.23. `books` and `01_20` intercorrelation also declined to 0.51.

Such big shift in just 3 out of 100 pairs tells us that our translated version of WordSim isn't robust enough, and aposterior tweaks are dangerous, because outcome can be easily manipulated.

For this reason correlation with translated WordSim cannot be compared to state-of-the-art values.

4.3 Corpus issues

Another interesting observation deals with over- and under-estimation of semantic relatedness.

Let the value of word pair relatedness be *underestimated* if its standard value is more than 3 points higher, than generated by algorythm, and *overestimated* if it is more than 3 points lower.

#	**Pair**		Relatedness estimation
1	Navalny (Навальный)	Putin (путин)	7.21
2	Navalny (Навальный)	power (власть)	7.18
3	Navalny (Навальный)	people (народ)	7.07
4	Navalny (Навальный)	president (президент)	7.05
1	iPhone (айфон)	telephone (телефон)	7.63
2	iPhone (айфон)	computer (комп)	7.17
3	iPhone (айфон)	internet (интернет)	7.09
1	internet (интернет)	work (работать)	8.44
2	internet (интернет)	problem (проблема)	8.41
3	internet (интернет)	inet (inet)	8.36

Table 3: Terms defined by Twitter corpus

It turns out that `books` underestimate value in 45% cases, and never overestimate; `01_20` overestimate value in 35% cases, and never underestimate. They both give large estimation error in only 2% cases, i.e. they guess wrong in different cases.

4.4 Empirical study

We also made several empirical studies. Taking into account that Twitter posts are often related to current events and trends, we assume that it is possible to identify these events and get some knowledge of what do they look like.

We take three words and display their most related (according to `01_20`) alies. They can be seen at Table 3.

The first one is the name of Russian oppositioner, A. Navalny. Words most related to him are *Putin*, *power*, *people* and *president*, which seen to be related to his public image.

The last one, *internet*, is actually more blurred, and seem unrelated, but it can be explained that people often post Tweets about problems with internet link quality.

5 Discussion

During this research we evaluated the translated version of WordSim353. It turned out that such method can be used to estimate accuracy of semantic relatedness algorythm.

Although term *accuracy* seems a little bit confusing, because it cannot be compared to that of state-of-the-art (the absolute value). It still can be used as relative quality measure, e.g. for estimating quality change between two versions of same algorythm.

We also evaluated Twitter stream as linguistic resource, which performed slightly worse than same algorythm trained with set of several Russian contemporary literature texts. However, Twitter-based dataset was an order of magnitude larger than former one.

Twitter stream can also be used to estimate meaning of new words. It also may be considered to detect word semantic change.

6 Future work

There are number of ways to continue this work.

WordSim353 can be translated completely and analogous evaluation performed. As well as other standards.

Several other resources may be tried, e.g. Wikipedia articles, Wictionary and WordNets.

Several modifications may be applied to the algorythm, including different `td-idf` modificatons, different stopword lists, and usage of methods introduced by (Mikolov et al., 2013).

Twitter-based semantic relatedness may be combined with trend analysis to produce trending topics.

References

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*, pages 165–171. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. *LREC*, 8:1646–1652.

Appendix A. Texts in `books` corpus.

B. Akunin. Almaznaya kolesnitsa.

B. Akunin. Vneklassnoe chtenie.

D. Granin. Zubr.

V. Pelevin. Prince gosplana.

V. Pelevin. Pokolenie ”P”.

S. Lukianenko. 13 gorod.

S. Lukianenko. Pristan zheltih korablei.

A. Strugatsky, B. Strugatsky. Ponedelnik nachinaetsa v subbotu.

A. Strugatsky, B. Strugatsky. Ulitka na sklone.

M. Veller. Vse o zhizni.

A. Zhitinksi. Ditia epokhi.