Alternative measures of word relatedness in distributional semantics

Alina Maria Ciobanu
Faculty of Mathematics
and Computer Science
University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro

Anca Dinu
Faculty of Foreign Languages
and Literatures
University of Bucharest
anca_d_dinu@yahoo.com

Abstract
This paper presents an alternative method to measuring word-word semantic relatedness in distributional semantics framework. The main idea is to represent target words as rankings of all co-occurring words in a text corpus, ordered in their \( tf - idf \) weight and use a metric between rankings (such as Jaro distance or rank distance) to compute semantic relatedness. This method has several advantages over the standard approach that uses cosine measure in a vector space, mainly in that it is computationally less expensive (i.e. does not require working in a high dimensional space, employing only rankings and a distance which is linear in the rank’s length) and presumably more robust. We tested this method on the standard WS-353 Test, obtaining the co-occurrence frequency from the Wacky corpus. The results are comparable to the methods which use vector space models; and, most importantly, the method can be extended to the very challenging task of measuring phrase semantic relatedness.

1 Introduction
This paper presents a method of measuring word-word semantic relatedness in the distributional semantics (DS) framework.

DS relies on a usage-based perspective on meaning, assuming that the statistical distribution of words in context plays a key role in characterizing their semantic behavior. The idea that word co-occurrence statistics extracted from text corpora can provide a basis for semantic representations can be traced back at least to (Firth, 1957): “You shall know a word by the company it keeps” and (Harris, 1954) "words that occur in similar contexts tend to have similar meanings”. This view is complementary to the formal semantics perspective, focusing on the meaning of content words, (such as nouns, adjectives, verbs or adverbs) and not on grammatical words (prepositions, auxiliary verbs, pronouns, quantifiers, coordination, negation), which are the focus of formal semantics. Since many semantic issues come from the lexicon of content words and not from grammatical terms, DS offers semantical insight into problems that cannot be addressed by formal semantics.

Moreover, DS Models can be induced fully automatically on a large scale, from corpus data. Thus, a word may be represented by a vector in which the elements are derived from the occurrences of the word in various contexts, such as windows of words (Lund and Burgess, 1996), grammatical dependencies (Dekang, 1998; Padó and Lapata, 2007), and richer contexts consisting of dependency links and selectional preferences on the argument positions (Erk and Padó, 2008).

The task of measuring word-word relatedness was previously performed in DS by using vector space models (see (Turney and Pantel, 2010) for an excellent survey of vector-space models), that is employing high dimensional matrices to store co-occurrence frequency of target words and some set of dimension words, usually highly frequent (but not grammatical) words. The relatedness of two target words was typically given by the cosine of the angle between their vectors. Instead of using vector space models, we propose to represent the target words only by rankings (vectors) of words in their decreasing order of co-occurrence frequency or their \( tf - idf \) weight. The \( tf - idf \) weight increases with the number of co-occurrences and with the "selectiveness" of the term - the fewer distinct words it occurs with, the higher the weight.

This proposal has some advantages, as discussed in Approach section. We can measure the semantic relatedness between two target words by computing the distance between the two cor-
responding rankings, using distances defined on rankings.

In the remaining of the paper we will present our approach, describe the data we have used, compare the results and draw the conclusions.

2 Approach

The method we propose is meant to measure word-word semantic relatedness, in a bag of words model, using 4 different distances (Rank Distance, MeanRank Distance, CosRank Distance and Jaro Distance) between rankings. To do so, instead of representing words in vector spaces, we represent them as rankings of co-occurring words ordered after their semantic contribution, i.e. arranged in their raw co-occurrence frequency and, separately, in their tf-idf weight. We thus take into consideration all words that co-occurred with a target word, not just a predefined set of dimension words.

We define the rank distance (variants) and the Jaro distance, as it follows.

A ranking is an ordered list and is the result of applying an ordering criterion to a set of objects. Formally (Dinu, 2005), we have:

Let $U = \{1, 2, ..., \#U\}$ be a finite set of objects, named universe (we write $\#U$ for the cardinality of $U$). A ranking over $U$ is an ordered list: $\tau = (x_1 > x_2 > ... > x_d)$, where $x_i \in U$ for all $1 \leq i \leq d$, $x_i \neq x_j$ for all $1 \leq i \neq j \leq d$, and $>$ is a strict ordering relation on the set $\{x_1, x_2, ..., x_d\}$.

A ranking defines a partial function on $U$ where for each object $i \in U$, $\tau(i)$ represents the position of the object $i$ in the ranking $\tau$.

The order of an object $x \in U$ in a ranking $\sigma$ of length $d$ is defined by $ord(\sigma, x) = |d + 1 - \sigma(x)|$. By convention, if $x \not\in U \setminus \sigma$, then $ord(\sigma, x) = 0$.

Given two partial rankings $\sigma$ and $\tau$ over the same universe $U$, the rank distance between them is defined as:

$$\Delta(\sigma, \tau) = \sum_{x \in U \setminus \tau} |ord(\sigma, x) - ord(\tau, x)|.$$

MeanRank distance computes the average value of Rank distance computed when elements are ranked top-down and Rank distance computed when elements are ranked bottom-up.

Given two full rankings $\sigma$ and $\tau$ over the same universe $U$ with $|U| = n$, CosRank distance (Dinu and Ionescu, 2012) is defined as follows:

$$\Delta(\sigma, \tau) = \frac{\sum_{x \in U} ord(x|\sigma) \times ord(x|\tau)}{\sum_{x \in U} |ord(x|\sigma)|^2 + |ord(x|\tau)|^2 + ... + n^2}.$$
2.1 The data
We used the publicly available Wacky corpus (Baroni et al., 2009). The corpus is lemmatized and pos tagged. As it is usual in distributional semantics, we only targeted content words and not grammatical words. Here is the list with the pos tags we have employed:

- JJ adjective, e.g. green
- JJR adjective, comparative, e.g. greener
- JJS adjective, superlative, e.g. greenest
- NN noun, singular or mass, e.g. table
- NNS noun plural, e.g. tables
- NPS proper noun, plural, e.g. Vikings
- RB adverb, e.g. however, usually, naturally, here, good
- VV verb, base form, e.g. take
- VVD verb, past tense, e.g. took
- VVG verb, gerund/present participle, e.g. taking
- VVN verb, past participle, e.g. taken
- VVP verb, sing. present, non-3d, e.g. take
- VVZ verb, 3rd person sing., present, e.g. takes

Accordingly, we have extracted from Wacky corpus the 10 words window co-occurrence vectors for the words in WS-353 Test (Finkelstein et al., 2002). WS-353 Test is a semantic relatedness test set consisting of 353 word pairs and a gold standard defined as the mean value of evaluations by up to 17 human judges. The value scale for the test is from 0 to 10: completely unrelated words were assigned a value of 0, while identical words a value of 10. Although this test suite contains some controversial word pairs, and there are other test suites such as in (Miller and Charles, 1991) and (Rubenstein and Goodenough, 1965), it has been widely used in the literature and has become the de facto standard for semantic relatedness measure evaluation. For all the 437 target-words in WS-353 Test, we computed the raw co-occurrence frequency $tf_{t,d}$ of terms $t$ (base-word) and $d$ (target-word), defined as the number of times that $t$ and $d$ co-occurred. We preprocessed the data, as it follows:

- we deleted all non-English words;
- we eliminated all other words containing non-letter characters;
- we separated hyphenated words and recomputed the weight accordingly;

Then we standardly processed the raw co-occurrence frequencies, transforming it into the $tf - idf$ weight: $w_{t,d} = (1 + \log_{10}tf_{t,d}) \times \log_{10}N/df_t$, where $N = 437$ (the total number of words we are computing vectors for) and $df_t$ is the number of target words $t$ co-occurs with. The $tf - idf$ weight increases with the number of co-occurrences of $t$ and $d$ (co-occurrence frequency) and increases with the ”selectiveness” of the term - the fewer distinct words it occurs with, the higher the weight.

We then computed the distances between pairs of target words both for raw frequencies and for tf-idf weights, for different lengths of the rankings, starting with a length of only 10 and adding 10 at a time until 2000.

3 Results
We summarize our results in Figure 1: one graphic for experiments with raw frequencies and one for experiments with tf-idf weight. On the OX axis we represent the length of the rankings (up to the first 2000 words) and on the OY axis the value of human/machine correlation. We only represented the best 3 performing distances, namely Rank, CosRank and Jaro, along with the standard Cosine Distance (for comparison).

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>Spearman correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hughes and Ramage (2007)</td>
<td>WordNet</td>
<td>0.55</td>
</tr>
<tr>
<td>Finkelstein et al. (2002)</td>
<td>LSA, Combination</td>
<td>0.56</td>
</tr>
<tr>
<td>Gabriolovich and Markovitch (2007)</td>
<td>ODP</td>
<td>0.65</td>
</tr>
<tr>
<td>Agirre et al. (2009)</td>
<td>Web Corpus</td>
<td>0.65</td>
</tr>
<tr>
<td>Agirre et al. (2010)</td>
<td>WordNet</td>
<td>0.69</td>
</tr>
<tr>
<td>Gabriolovich and Markovitch (2007)</td>
<td>Wikipedia</td>
<td>0.75</td>
</tr>
<tr>
<td>Agirre et al. (2009)</td>
<td>Combination</td>
<td>0.78</td>
</tr>
<tr>
<td>This work</td>
<td>Wacky</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 1: Comparison with vector space experiments for WordSim353.

For the raw co-occurrence, one observe that until the lengths of 1000, the best performing distance was Jaro distance, followed by CosRank, Rank, all three of them outperforming Cosine. Between a length of 1000 and 2000, the order reverses and Cosine is the best performing distance. An explanation for this is on the one hand that Jaro and Rank distance need no preprocessing like computing $tf - idf$ weight and, on the other, that words ranked on places over a certain threshold (in this case 1000) are, in fact, irrelevant (or even represent noise) for the semantic representation of the target word. For the $tf - idf$ weight, the traditional Cosine distance performs best, while CosRank is on the second place.
Overall, it turns out that the differences are minor and that measuring the distances between rankings instead of vectors is a valid option. The results may thus be further used as baseline for experimenting with this method, like, for instance taking syntactic structure into account.

As we can see in Table 1, the best correlation value of 0.55 (obtained by CosRank computed on the tf-idf weights) is identical to the baseline correlation values for the vector space experiments.

When inspecting the worst mismatches between human/machine relatedness judgments between pairs of words, we observed that most of them were following a pattern, namely lower values assigned by humans almost always corresponded to much higher values computed by machine, such in the following examples given in Table 2:

<table>
<thead>
<tr>
<th>Word Pair</th>
<th>Human distance</th>
<th>Machine Distance (Jaro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(month, hotel)</td>
<td>1.81</td>
<td>6,293567</td>
</tr>
<tr>
<td>(money, operation)</td>
<td>3.31</td>
<td>6,40989</td>
</tr>
<tr>
<td>(king, cabbage)</td>
<td>0.23</td>
<td>4,171145</td>
</tr>
<tr>
<td>(coast, forest)</td>
<td>3.15</td>
<td>6,409761</td>
</tr>
<tr>
<td>(rooster, voyage)</td>
<td>0.62</td>
<td>4,656631</td>
</tr>
<tr>
<td>(governor, interview)</td>
<td>3.25</td>
<td>6,08319</td>
</tr>
<tr>
<td>(drink, car)</td>
<td>3.04</td>
<td>5,931482</td>
</tr>
<tr>
<td>(day, summer)</td>
<td>3.94</td>
<td>6,576498</td>
</tr>
<tr>
<td>(architecture, century)</td>
<td>3.78</td>
<td>5,927582</td>
</tr>
<tr>
<td>(morality, marriage)</td>
<td>3.69</td>
<td>5,450308</td>
</tr>
</tbody>
</table>

Table 2: Comparison with vector space experiments for WordSim353.

One can intuitively speculate about the reason of these differences; for instance, the pairs (summer day) and (king, cabbage) are present in the data as collocations: "summer day" and "king cabbage", which is a very large variety of cabbage. The other pairs (month, hotel), (money, operation), (rooster, voyage), etc. seem to allow for explanations based on pragmatic information present in the data.

4 Conclusions and further work

We introduced in this paper an alternative method to measuring word-word semantic relatedness; instead of using vector space models, we proposed to represent the target words only by rankings (vectors) of words in their decreasing order of co-occurrence frequency; we computed the word-word relatedness by four different distances. We tested this method on the standard WS-353 Test, obtaining the co-occurrence frequency from the Wacky corpus. The Spearman correlation with human given scores are around the baseline for vector space models, so there is hope for improvement. The method is computationally less expensive. Furthermore, it provides a new framework for experimenting with distributional semantic compositionality, since our method can be extended from measuring word-word semantic relatedness to evaluating phrasal semantics. This is in fact one of the most challenging streams of research on distributional semantics: finding a principled way to account for natural language compositionality.

In the future, we will extend the contribution in this paper to evaluating phrase semantics, that differs from all the above methods in that it does not try to learn weights or functions for the vectors, but instead combines or aggregates two vectors containing words ranked in their semantic contribution, in order to obtain a vector for the resulting phrase. When combining two word vectors, one obtains an aggregation set which contains all vectors for which the sum of the distances between them and the two vectors is minimum. The vector in the aggregation set that is closest to the syntactic head of the new phrase is chosen to be the vector representing it. Thus, the syntactic structure of the phrase is taken into account. The word - phrase semantic similarity can be computed as in the experiment reported in this paper and the obtained values compared to some golden standard, like, for instance, in SemEval 2013 task, Evaluating Phrasal Semantics or like the dataset in (Mitchell and Lapata, 2008).
Acknowledgments

This work was supported by the research project PN-II-ID-PCE-2011-3-0959.

References


