Distributional Similarity of Words with Different Frequencies

Christian Wartena
Hochschule Hannover, University of Applied Sciences and Arts
Expo Plaza 12
30359 Hannover, Germany
christian.wartena@hs-hannover.de

ABSTRACT

Distributional semantics tries to characterize the meaning of words by the contexts in which they occur. Similarity of words hence can be derived from the similarity of contexts. Contexts of a word are usually vectors of words appearing near to that word in a corpus. It was observed in previous research that similarity measures for the context vectors of two words depend on the frequency of these words. In the present paper we investigate this dependency in more detail for one similarity measure, the Jensen-Shannon divergence. We give an empirical model of this dependency and propose the deviation of the observed Jensen-Shannon divergence from the divergence expected on the basis of the frequencies of the words as an alternative similarity measure. We show that this new similarity measure is superior to both the Jensen-Shannon divergence and the cosine similarity in a task, in which pairs of words, taken from Wordnet, have to be classified as being synonyms or not.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods, Linguistic Processing*; G.3 [**Probability and Statistics**]: [Correlation and regression analysis]; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Models,Text Analysis*

General Terms

Experimentation

Keywords

Distributional Similarity, Synonymy

1. INTRODUCTION

For many applications dealing with texts it is useful or necessary to know what words in a language are similar. Similarity between words can be found in hand crafted resources, like WordNet [8], but methods to derive word similarities from large text corpora are at least an interesting alternative. Intuitively, words that occur in the same texts or, more generally, the same contexts are similar. Thus we could base a similarity measure on the number of times two words occur in the same context, e.g. by representing words in a document space. Especially, if we consider small contexts, like a window of a few words around a word, this approach gives pairs of words that are in some dependence relation to each other. De Saussure [3] calls such such relations, defined by co-presence in a linguistic structure (e.g. a text, sentence, phrase, fixed window, words in a certain grammatical relation to the studied word and so on), *syntagmatic* relations. The type of similarity that is much closer to synonymy and much more determined by the meaning of a word, is obtained by comparing the contexts in which a word occurs. This type of similarity is usually called *paradigmatic* similarity or distributional similarity.

Though distributional similarity has widely been studied and has established as a method to find similar words, there is no consensus on the way the context of a word has to be defined and on the best way to compute the similarity between contexts. In the most general definitions the context of a word consists of words and their relation to the given word (see e.g. [6, 2]). In the following we will only consider the simplest case in which there is only one relation: the relation of being in the same sentence. Now each word can be represented by a *context vector* in a high dimensional word space. Since these context vectors are very sparse, often dimensionality reduction techniques are applied. In the present paper we use random indexing, introduced by Karlgren and Sahlgren [7] and Sahlgren [9] to reduce the size of the context vectors. For random indexing each word is represented by a random index vector. The context vector of a word is constructed by addition of the index vectors of all words in the context. Thus the dimensionality of the context vector is the same as the dimensionality chosen for the index vectors. It was shown by Karlgren and Sahlgren [7] that this technique gives results that are comparable to those obtained by dimensionality reduction techniques like singular value decomposition, but requires less computational resources. The similarity of the context vectors, finally, can be used as a proxy for the similarity of words.

In order to evaluate the various methods to define context vectors and the various similarity measures that can be used subsequently, usually the computed similarity of words is tested in a task in which words have to be classified as being synonym or not to a given word. Often the data are taken from the synonym detection task from TOEFL (Test of English as a Foreign Language) in which the closest related word from a set of four words has to be chosen. Görnerup and Karlgren [5] found that best results are obtained using L1-norm or Jensen-Shannon divergence (JSD).

DIR 2013, April 26, 2013, Delft, The Netherlands.

Curran and Moens [2] obtain best results using a combination of the Jaccard coefficient and the T-test while Van der Plas and Bouma [10] report best results using a combination of the Dice coefficient and pointwise mutual information. Both Curran and Moens and Van der Plas and Bouma use a number of different relations and need a similarity measure that is able to assign different weights to the relations. This makes their results less relevant for the present paper. The differences between the latter two studies show how strongly the results depend on the exact settings of the experiment. Many authors, however, use cosine similarity as a generally well established similarity measure for vectors in high dimensional word spaces.

Weeds et al. [13] do not compare similarity measures to hand crafted data sets but studied characteristic properties of various measures. They find that, in a task where words related to a given word have to be found, some similarity measures tend to find words with a similar frequency as the target word, while others favor highly frequent words. The Jensen-Shannon divergence (JSD) is one of the measures that tends to favor more general terms. In the following we will investigate this in more detail. We show that a better similarity measure can be defined on the base of the JSD, when we use our knowledge about the dependency of the JSD on the frequency of the words. Finally, we show that this new similarity measure outperforms the original JSD and the cosine similarity in a task in which a large number of word pairs have to be classified as synonyms or non-synonyms.

2. INFLUENCE OF WORD FREQUENCY

As already mentioned above, Weeds et al. [13] observed that, in tasks in which related words have to be found, some measures prefer words with a frequency similar to that of the target word while others prefer highly frequent words, regardless of the frequency of the target word. The JSD belongs to the latter category. In Wartena et al. [12] we also made this observation. There we compared context vectors of words with the word distribution of a document with the goal of finding keywords for the document. In order to compensate for the strong bias to highly frequent words, we introduced specificity as an explicit second condition for finding keywords. As long as we try to find synonyms for a given word, i.e. if we compare pairs of words in which one component is fixed, like in the TOEFL tests, the problem usually is tolerable. Moreover, the problem is not that apparent if the range of the lowest and highest frequencies is not too large, e.g. when only words with certain minimal frequency are considered and the size of the corpus gives a low upper bound on the frequency. Length effects are completely avoided if for every word the same amount of contexts is sampled, as e.g. is done by Giesbrecht [4]. As we will see below, JSD becomes completely useless if we compare arbitrary word pairs and do not pose any lower or upper bound on the frequency of the words.

The JSD between two probability distributions is defined as the average of the relative entropy of each of the distributions to their average distribution. It is interesting to note, that the JSD can be written as

$$\begin{aligned}\mathrm{JSD}(p,q) &= \tfrac{1}{2}D(p||\tfrac{1}{2}p+\tfrac{1}{2}q)+\tfrac{1}{2}D(q||\tfrac{1}{2}p+\tfrac{1}{2}q)\\ &= \log 2 + \frac{1}{2}\sum_{t:p(t)\neq 0\,\wedge\, q(t)\neq 0}\Big(p(t)\log\big(\frac{p(t)}{p(t)+q(t)}\big)\\ &\quad +q(t)\log\big(\frac{q(t)}{p(t)+q(t)}\big)\Big). \qquad (1)\end{aligned}$$

This formulation of the JSD explicitly shows that the value only depends on the words that have a non-zero value in both context vectors. If there is no common word the JSD is maximal. Now suppose that all words are independent. If the context vectors are based on a few instances of a word, the probability that a context word co-occurs with both words is rather low. To be a bit more precise, if we have context vectors v_1 and v_2 that are distributions over d elements, with n_1 and n_2 non zero elements, than the probability that a word is not zero in both distributions is, as a first approximation, $\frac{n_1}{d}\cdot\frac{n_2}{d}$. Even if the words are not independent, we might expect a similar behavior: the probability that a word has a non zero value in two context vectors increases with the number contexts on which the vectors are based.

If we try to predict the JSD of the context vectors of two words, we could base this prediction on the frequency of the words. However, it turns out that this is a very complicated dependency. Alternatively, we could base the prediction on the entropy of the context vector (if we interpret the vector as a probability distribution, as we have to do to compute the JSD): if the entropy of both vectors is maximal, they have to be identical and the JSD will be 0. If the entropy of both vectors is minimal, the JSD of the two vector is most likely to be maximal. Since, in case of independence of all words, the context vectors will not converge to the equal distribution but to the background distribution, i.e. the word distribution of the whole corpus, it is more natural to use the relative entropy to the background distribution. Preliminary experiments have shown that this works, but that JSD of two context vectors can be better predicted by the number of non-zero values in the vectors.

Figure 1 shows the relation between the JSD of two context vectors and the product of the number of non zero values in both distributions. The divergences in this figure are computed for distributions over 20 000 random indices computed on the 2,2 billion words ukWaC Corpus for 9916 word pairs. We found the same dependency for the L1 norm. In contrast, for the cosine similarity we could not find any dependency between the number of instances of the words or the number of non zero values in the context distributions.

3. EXPERIMENTAL RESULTS

To test our hypothesis that the divergence of two context vectors depends on the number of instances on which these vectors are based, we computed divergences for almost 10 000 word pairs on a very large corpus. Furthermore, we show how the knowledge about this dependency can be used to find a better measure to capture the semantic similarity between two words.

3.1 Data

As a corpus to compute the context distribution we use the POS tagged and lemmatized version of the ukWaC Corpus of approximately 2,2 billion words [1]. As the context of a

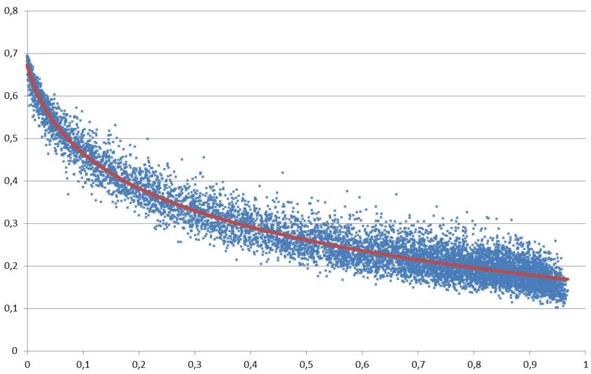

Figure 1: Divergence vs. product of relative number of non-zero values for pairs of context vectors and a function modeling the dependency.

word we consider all lemmata of open class words (i.e. nouns, adjectives, verbs, etc.) in the same sentence. We define a sentence simply as a set of words. A corpus then is a set of sentences. Let C be a corpus and w a word, then we define $C_w = \{S \in C \mid w \in S\}$. Given a corpus C, the context vector p_w of a word w can be defined as

$$p_w = \frac{1}{|C_w|} \sum_{S \in C_w} \frac{1}{|S|} \sum_{v \in S} r_v \tag{2}$$

where r_v is the random index vector of the word v. The random index vector is defined as a probability distribution over d elements, such that for some small set of random numbers $R = \{r \in \mathbb{N} \mid r < d\}$ there are n elements $r_v(i) = \frac{1}{|R|}$ if $i \in R$ and $r_v(i) = 0$ otherwise. In the following we will use distributions with $d = 20\,000$ and $|R| = 8$ unless stated else. Note, that we will always use probability distributions, but stick to the usual terminology of (context) vectors.

For the evaluation of the similarity measures we selected pairs of words from Wordnet [8]. We started with a list of pairs (w_1, w_2) such that (1) w_1 and w_2 are single words, (2) w_1 occurs at least two times in the British National Corpus and (3) w_1 and w_2 share at least one sense. This resulted in a list of 24 576 word pairs. From this list we selected all pairs for which the Jaccard coefficient of the sets of senses of the words is at least 0.7. After filtering out all pairs containing a word that was not found in the ukWaC corpus a list of 849 pairs remained. These word pairs are considered as synonyms in the following. Next from the list of 24 576 word pairs the second components were reordered randomly. The resulting list of new word pairs was filtered such that the two words of each pair both occur in the ukWaC corpus and have no common sense. This resulted in a list of 8967 word pairs.[1]

As a consequence of the requirement of the overlap of Wordnet senses, most words in the synonym list have very few senses and are very infrequent words. Thus the average frequency in ukWaC of the synonyms is much lower than that of the words of the non-synonym list. The most frequent word (*use*) was found 4.57 million times in the ukWaC corpus; 117 words were only found once (e.g. *somersaulting*, *sakartvelo*).

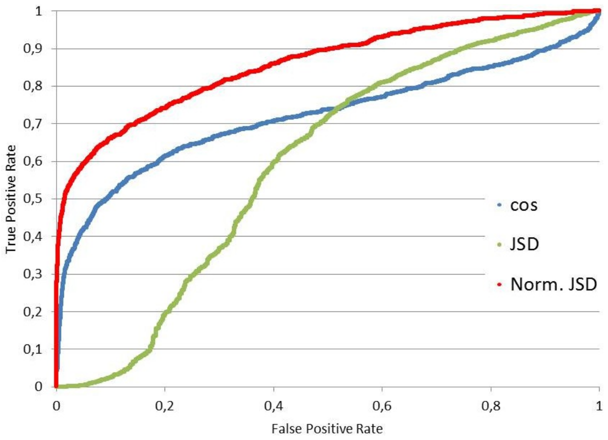

Figure 2: ROC Curves for ranking of word pairs (849 synonym pairs, 8967 non synonym pairs) using different similarity measures.

3.2 Predicting JSD of context vectors

Figure 1 shows that there is a clear dependency between the JSD of a word pair and the product of the relative numbers of non zero values in the context distributions. This dependency can be captured by the following equation:

$$\mathrm{JSD}^{\exp}(p_1, p_2) = a \log\left(1 + \frac{b}{n}\right) + c \tag{3}$$

with $n = \frac{n_1}{d} \cdot \frac{n_2}{d}$ where n_1 and n_2 are the number of non zero values of n_1 and n_2, respectively. Optimal values for a, b and c were found by maximizing the coefficient of determination, R^2, on all non-synonym word pairs. We left out the synonyms, since we try to model the similarity that is caused just by the probability of random words to occur in these context with an increasing number of observations. With $a = -0.34$, $b = 0.032$ and $c = 0.67$ a R^2 score of 0.95 is reached (0.93 for the same constants when synonyms are included). The curve corresponding to these values is displayed in red in Figure 1. Since usually context vectors with much less dimensions are used, we repeated the experiment with context distributions over 1 000 random indices and obtained a R^2 value of $0,92$ ($a = -1.65$, $b = 0.99$ and $c = 0.61$).

3.3 Ranking word pairs

Most of the variance in the JSD of two context distributions can be explained by (3). Now we expect that the remaining variance reflects the degree to which the words have a similar function or even meaning. To test this we define the *(frequency) normalized JSD* as

$$\mathrm{JSD}^{\mathrm{norm}}(p_1, p_2) = \mathrm{JSD}(p_1, p_2) - \mathrm{JSD}^{\exp}(p_1, p_2) \tag{4}$$

Ideally, all word pairs of synonyms will be ranked higher than the non-synonym pairs. We use the area under the ROC curve (AUC) to evaluate the ranking. We compare the ranking according to the normalized JSD with the rankings from the JSD, the cosine similarity and the L1 norm that is used sometimes in combination with random indexing. The L1 norm between two vectors v_1 and v_2 of dimensionality d is defined as $\sum_{0 \leq i < d} |v_1(i) - v_2(i)|$. The ROC curves are given in Figure 2 when using context vectors with 20 000 dimensions. The AUC-values are summarized in Table 1, both for

[1] The lists of word pairs are available at `http://nbn-resolving.de/urn:nbn:de:bsz:960-opus-4077`.

Table 1: AUC of classifying wordpairs as synonyms using different numbers of dimensions and different similarity measures

Number of dimensions	Similarity Measure	AUC
1000	Cosine	0,53
1000	JSD	0,41
1000	$\mathrm{JSD}^{\mathrm{norm}}$	0,52
20000	Cosine	0,72
20000	JSD	0,41
20000	L1	0,42
20000	$\mathrm{JSD}^{\mathrm{norm}}$	**0,86**

the experiment using context distributions over 20 000 and 1 000 random indices.

We see that the JSD gives a ranking worse than a random ranking. The remarkable observation is the large difference between the AUC values, since we are comparing exactly the same context distributions, and thus use exactly the same information. A further observation is the strange behavior of the cosine similarity. For pairs of words for which less than a dozen instances were found, the cosine similarity seems to give almost random results. Thus some positive pairs are ranked very low, explaining the rise of the ROC curve at the right end. The results of the L1 norm are almost the same as those of the JSD, which is not surprising as we also found a linear correspondence between JSD and the L1 norm.

Finally, it should be noted that we did not try to find the best possible ranking. If we would include frequency information (two very frequent words are unlikely to be synonyms) or Levenshtein distance (there are many spelling variants included in the list of synonyms) we could easily obtain a better ranking. The goal of the experiment, however, was evaluation of distance measures for random indexing. The classification is only a means to assess the quality of the distance measure. In [11] we also investigate the possibility to combine various distance measures and other features to get an optimal ranking.

4. DISCUSSION AND CONCLUSIONS

We have clearly found a very strong dependency between the number of non-zero values in random context vectors and the JSD between the vectors. When we use data with an extremely large range in frequencies this leads to JSD values that are useless for ranking word pairs according to their similarity. Note that we included words with frequencies ranging from 1 to 4,57 Million. We used the known dependency between the number of non zero values in the distributions and the JSD to define a new similarity measure, the frequency normalized JSD. This measure clearly outperforms the cosine similarity in the ranking experiment.

Though this result is convincing, we are lacking a theoretical base from which a formula like (3) can be derived. Also, it would be preferable if the constants could be estimated directly from the size of the corpus, the number of dimensions, etc. Now, only one from three constants can easily be explained, namely as the maximum JSD. Alternatively, also smoothing of the context distributions might be a solution to make JSD more useful. The smoothing should then account for the similarities that stem from random words appearing in both contexts. In general, the results show that the choice for the right similarity measure to be used for distributional similarity is not a solved question and more research in this area is needed.

5. REFERENCES

[1] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43 (3): 209-226*, 43(3):209–226, 2009.

[2] J. R. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLAX).*, pages 59–66. Association of Computational Linguistics, 2002.

[3] F. de Saussure. *Cours de linguistique générale.* V.C. Bally and A. Sechehaye (eds.), Paris/Lausanne, 1916. English translation: Course in General Linguistics. London: Peter Owen, 1960.

[4] E. Giesbrecht. Towards a matrix-based distributional model of meaning. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 23–28, Los Angeles, California, 2010. ACL.

[5] O. Görnerup and J. Karlgren. Cross-lingual comparison between distributionally determined word similarity networks. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 48–54. ACL, 2010.

[6] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 89–97. ACM, 1992.

[7] J. Karlgren and M. Sahlgren. From words to understanding. In *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications, Stanford, Californa, 2001.

[8] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[9] M. Sahlgren. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5, 2005.

[10] L. Van Der Plas and G. Bouma. Syntactic contexts for finding semantically related words. In *Proceedings of Computational Linguistics in the Netherlands*, 2004.

[11] C. Wartena. HsH: Estimating semantic similarity of words and short phrases with frequency normalized distance measures. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 2013. to appear.

[12] C. Wartena, R. Brussee, and W. Slakhorst. Keyword extraction using word co-occurrence. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*, pages 54–58. IEEE, 2010.

[13] J. Weeds, D. J. Weir, and D. McCarthy. Characterising measures of lexical distributional similarity. In *COLING 2004, Proceedings of the 20th International Conference on Computational Linguistics*, 2004.