Feature Vector Quality and Distributional Similarity

Maayan Geffet
School of Computer Science and Engineering,
Hebrew University
Givat Ram Campus,
Jerusalem, Israel, 91904
mary@cs.huji.ac.il

Ido Dagan
Department of Computer Science,
Bar-Ilan University
Ramat-Gan, Israel, 52900

dagan@cs.biu.ac.il

Abstract

We suggest a new goal and evaluation criterion for word similarity measures. The new criterion - *meaning-entailing substitutability* - fits the needs of semantic-oriented NLP applications and can be evaluated directly (independent of an application) at a good level of human agreement. Motivated by this semantic criterion we analyze the empirical quality of distributional word feature vectors and its impact on word similarity results, proposing an objective measure for evaluating feature vector quality. Finally, a novel feature weighting and selection function is presented, which yields superior feature vectors and better word similarity performance.

1 Introduction

Distributional Similarity has been an active research area for more than a decade (Hindle, 1990), (Ruge, 1992), (Grefenstette, 1994), (Lee, 1997), (Lin, 1998), (Dagan et al., 1999), (Weeds and Weir, 2003). Inspired by Harris distributional hypothesis (Harris, 1968), similarity measures compare a pair of weighted feature vectors that characterize two words. Features typically correspond to other words that co-occur with the characterized word in the same context. It is then assumed that different words that occur within similar contexts are semantically similar.

As it turns out, distributional similarity captures a somewhat loose notion of semantic similarity (see Table 1). By construction, if two words are distributionally similar then the occurrence of one word in some contexts indicates that the other word is also likely to occur in such contexts. But it does not ensure that the meaning of the first word is preserved when replacing it with the other one in the given context. For example, words of similar semantic types, such as *company – government*, tend to come up as distributionally similar, even though they are not substitutable in a meaning preserving sense.

On the other hand, many semantic-oriented applications, such as Question Answering, Paraphrasing and Information Extraction, do need to recognize which words may substitute each other in a meaning preserving manner. For example, a question about *company* may be answered by a sentence about *firm*, but not about *government*. Such applications usually utilize reliable taxonomies or ontologies like WordNet, but cannot rely on the "loose" type of output of distributional similarity measures.

In recent work Dagan and Glickman (2004) observe that applications usually do not require a strict meaning preserving criterion between text expressions, but rather need to recognize that the meaning of one expression *entails* the other. Entailment modeling is thus proposed in their work as a generic (application-independent) framework for practical semantic inference. We suggest adopting such (directional) entailment criterion at the lexical level for judging whether one word can be substituted by another one. For example, certain questions about companies might be answered by sentences about automakers, since the meaning of *automaker* entails the meaning of *company* (though not vice versa). In this paper we adapt this new criterion, termed *meaning entailing substitutability*, as a direct evaluation criterion for the "correctness" of the output of word similarity measures (as opposed to indirect evaluations through WSD or distance in WordNet).

Our eventual research goal is improving word similarity measures to predict better the more delicate meaning entailment relationship between words. As a first step it was necessary to analyze the typical behavior of current similarity measures and categorize their errors (Section 3). Our main observation is that the quality of similarity scores

nation	1	*city	7	*north	13	*company	19
region	2	territory	8	*economy	14	*industry	20
state	3	area	9	*neighbor	15	kingdom	25
*world	4	*town	10	*member	16	european_country	31
island	5	republic	11	*party	17	place	36
province	6	african_country	12	*government	18	colony	41

Table 1: The 20 top most similar words of *country* (and their ranks) in the similarity list by Lin98, followed by the next 4 words in the similarity list that are judged as correct. Incorrect similarities, under the substitutability criterion, are marked with '*'.

is often hurt by improper feature weights, which yield rather noisy feature vectors. We quantify this problem by a new measure for feature vector quality, which is independent of any particular vector similarity measure.

To improve feature vector quality a novel feature weighting function is introduced, called *relative feature focus (RFF)* (Section 4). While having a simple (though non-standard) definition, this function yields improved performance relative to the two suggested evaluation criteria – for vector quality and for word similarity. The underlying idea is that a good characteristic feature for a word *w* should characterize also multiple words that are highly similar to *w*. In other words, such feature should have a substantial "focus" within the close semantic vicinity of *w*.

Applying *RFF* weighting achieved about 10% improvement in predicting meaning entailing substitutability (Section 5). Further analysis shows that *RFF* also leads to "cleaner" characteristic feature vectors, which may be useful for additional feature-based tasks like clustering.

2 Background and Definitions

In the distributional similarity scheme each word *w* is represented by a feature vector, where an entry in the vector corresponds to a feature *f*. Each feature represents another word (or term) with which *w* co-occurs, and possibly specifies also the syntactic relation between the two words. The value of each entry is determined by some weight function *weight(w,f)*, which quantifies the degree of statistical association between the feature and the corresponding word.

Typical feature weighting functions include the logarithm of the frequency of word-feature co-occurrence (Ruge, 1992), and the conditional probability of the feature given the word (within probabilistic-based measures) (Pereira et al., 1993), (Lee, 1997), (Dagan et al., 1999). Probably the most widely used association weight function is (point-wise) Mutual Information (MI) (Church et al., 1990), (Hindle, 1990), (Lin, 1998), (Dagan, 2000), defined by:

$$MI(w, f) = \log_2 \frac{P(w, f)}{P(w)P(f)}$$

A known weakness of MI is its tendency to assign high weights for rare features. Yet, similarity measures that utilize MI showed good performance. In particular, a common practice is to filter out features by minimal frequency and weight thresholds. A word's vector is then constructed from the remaining features, which we call here *active* features.

Once feature vectors have been constructed, the similarity between two words is defined by some vector similarity metric. Different metrics have been used in the above cited papers, such as Weighted Jaccard (Dagan, 2000), cosine (Ruge, 1992), various information theoretic measures (Lee, 1997), and others. We picked the widely cited and competitive (e.g. (Weeds and Weir, 2003)) measure of Lin (1998) as a representative case, and utilized it for our analysis and as a starting point for improvement.

2.1 Lin's ('98) Similarity Measure

Lin's similarity measure between two words, *w* and *v*, is defined as follows:

$$sim(w, v) = \frac{\sum_{f \in F(w) \cap F(v)} weight(w, f) + weight(v, f)}{\sum_{f \in F(w)} weight(w, f) + \sum_{f \in F(v)} weight(v, f)},$$

where *F(w)* and *F(v)* are the active features of the two words and the weight function is defined as *MI*. A feature is defined as a pair <*term*, *syntac-*

Feature	Weight
Commercial-bank, gen↓	8.08
Destination, pcomp↑	7.97
Airspace, pcomp ↓	7.83
Landlocked, mod ↑	7.79
Trade_balance, gen ↓	7.78
Sovereignty, pcomp↓	7.78
Ambition , nn ↓	7.77
Bourse, gen ↓	7.72
Politician, gen ↓	7.54
Border, pcomp ↓	7.53

Table 2: The top-10 ranking features for *country*.

tic_relation>. For example, given the word "company" the feature <*earnings_report, gen* ↓ > (genitive) corresponds to the phrase "company's earnings report", and <*profit, pcomp* ↓ > (prepositional complement) corresponds to "the profit of the company". The syntactic relations are generated by the Minipar dependency parser (Lin, 1993). The arrows indicate the direction of the syntactic dependency: a downward arrow indicates that the feature is the parent of the target word, and the upward arrow stands for the opposite.

In our implementation we filtered out features with overall frequency lower than 10 in the corpus and with *MI* weights lower than 4. (In the tuning experiments the filtered version showed 10% improvement in precision over no feature filtering.) From now on we refer to this implementation as *Lin98*.

3 Empirical Analysis of Lin98 and Vector Quality Measure

To gain better understanding of distributional similarity we first analyzed the empirical behavior of Lin98, as a representative case for state of the art (see Section 5.1 for corpus details).

As mentioned in the Introduction, distributional similarity may not correspond very tightly to meaning entailing substitutability. Under this judgment criterion two main types of errors occur: (1) word pairs that are of similar semantic types, but are not substitutable, like *firm* and *government*; and (2) word pairs that are of different semantic types, like *firm* and *contract*, which might (or might not) be related only at a topical level. Table 1 shows the top most similar words for the target

Country-State	Ranks		Country-Economy	Ranks	
Broadcast	24	50	Devastate	81	8
Goods	140	16	Developed	36	78
Civil_servant	64	54	Dependent	101	26
Bloc	30	77	Industrialized	49	85
Nonaligned	55	60	Shattered	16	141
Neighboring	15	165	Club	155	38
Statistic	165	43	Black	122	109
Border	10	247	Million	31	245
Northwest	41	174	Electricity	130	154

Table 3: The top-10 common features for the word pairs *country-state* and *country-economy*, along with their corresponding ranks in the sorted feature lists of the two words.

word *country* according to Lin98. The two error types are easily recognized, e.g. *world* and *city* for the first type, and *economy* for the second.

A deeper look at the word feature vectors reveals typical reasons for such errors. In many cases, high ranking features in a word vector, when sorting the features by their weight, do not seem very characteristic for the word meaning. This is demonstrated in Table 2, which shows the top-10 features in the vector of *country*. As can be seen, some of the top features are either too specific (*landlocked*, *airspace*), and so are less reliable, or too general (*destination*, *ambition*), and hence not indicative and may co-occur with many different types of words. On the other hand, more characteristic features, like *population* and *governor*, occur further down the list, at positions 461 and 832. Overall, features that characterize well the word meaning are scattered across the ranked list, while many non-indicative features receive high weights. This may yield high similarity scores for less similar word pairs, while missing other correct similarities.

An objective indication of the problematic feature ranking is revealed by examining the common features that contribute mostly to the similarity score of a pair of similar words. We look at the common features of the two words and sort them by the sum of their weights in the two word vectors (which is the enumerator of Lin's *sim* formula in Section 2.1). Table 3 shows the top-10 common features for a pair of substitutable words (*country - state*) and non-substitutable words (*country - economy*). In both cases the common features are scattered across each feature vector, making it difficult

to distinguish between similar and non-similar word pairs.

We suggest that the desired behavior of feature ranking is that the common features of truly similar words will be concentrated at the top ranks of their vectors. The common features for non-similar words are expected to be scattered all across each of the vectors. More formally, given a pair of similar words (judged as substitutable) *w* and *v* we define the *top joint feature rank* criterion for evaluating feature vector quality:

$$top-rank(w,v,n) = \frac{1}{n}\sum_{f \in top-n(F(w) \cap F(v))} \frac{1}{2}[rank(w,f) + rank(v,f)],$$

where *rank(w,f)* is the feature's position in the sorted vector of the word *w*, and *n* is the number of top joint features to consider (*top-n*), when sorted by the sum of their weights in the two word vectors. We thus expect that a good weighting function would yield (on average) a low *top-rank* score for truly similar words.

4 Relative Feature Focus (RFF)

Motivated by the observations above we propose a new feature weight function, called *relative feature focus (RFF)*. The basic idea is to promote features which characterize many words that are highly similar to *w*. These features are considered as having a strong "focus" around *w*'s meaning. Features which do not characterize sufficiently many words that are sufficiently similar to *w* are demoted. Even if such features happen to have a strong direct association with *w* they are not considered reliable, as they do not have sufficient statistical mass in *w*'s semantic vicinity.

4.1 RFF Definition

RFF is defined as follows. First, a standard word similarity measure *sim* is computed to obtain initial approximation of the similarity space (Lin98 was used in this work). Then, we define the *word set* of a feature *f*, denoted by *WS(f)*, as the set of words for which *f* is an active feature. The semantic *neighborhood* of *w*, denoted by *N(w)*, is defined as the set of all words *v* which are considered sufficiently similar to *w*, satisfying $sim(w,v) > s$ where *s* is a threshold (0.04 in our experiments). *RFF* is then defined by:

$$RFF(w,f) = \sum_{v \in WS(f) \cap N(w)} sim(w,v).$$

That is, we identify all words *v* that are in the semantic neighborhood of *w* and are also characterized by *f* and sum their similarities to *w*.

Notice that *RFF* is a sum of word similarity values rather than being a direct function of word-feature association values (which is the more common approach). It thus does not depend on the exact co-occurrence level between *w* and *f*. Instead, it depends on a more global assessment of the association between *f* and the semantic vicinity of *w*. Unlike the entropy measure, used in (Grefenstette, 1994), our "focused" global view is *relative* to each individual word *w* and is not a global independent function of the feature.

We notice that summing the above similarity values captures simultaneously a desired balance between feature specificity and generality, addressing the observations in Section 3. Some features might characterize just a single word that is very similar to *w*. But then the sum of similarities will include a single element, yielding a relatively low weight.[1] General features may characterize more words within *N(f)*, but then on average the similarity with *w* over multiple words is likely to become lower, contributing smaller values to the sum. A reliable feature has to characterize multiple words (not too specific) that are highly similar to *w* (not too general).

4.2 Re-computing Similarities

Once *RFF* weights have been computed they are sufficiently accurate to allow for aggressive feature reduction. In our experiments it sufficed to use only the top 100 features for each word in order to obtain optimal results, since the most informative features now have the highest weights. Similarity between words is then recomputed over the reduced vectors using Lin's *sim* function (in Section 2.1), with *RFF* replacing *MI* as the new *weight* function.

[1] This is why the sum of similarities is used rather than an average value, which might become too high by chance when computed over just a single element (or very few elements).

5 Evaluation

5.1 Experimental Setting

The performance of the *RFF*-based similarity measure was evaluated for a sample of nouns and compared with that of Lin98. The experiment was conducted using an 18 million tokens subset of the Reuters RCV1 corpus,[2] parsed by Lin's Minipar dependency parser (Lin, 1993). We considered first an evaluation based on WordNet data as a gold standard, as in (Lin, 1998; Weeds and Weir, 2003). However, we found that many word pairs from the Reuters Corpus that are clearly substitutable are not linked appropriately in WordNet.

We therefore conducted a manual evaluation based on the judgments of two human subjects. The judgment criterion follows common evaluations of paraphrase acquisition (Lin and Pantel, 2001), (Barzilay and McKeown, 2001), and corresponds to the meaning-entailing substitutability criterion discussed in Section 1. Two words are judged as *substitutable* (correct similarity) if there are some contexts in which one of the words can be substituted by the other, such that the meaning of the original word can be inferred from the new one.

Typically substitutability corresponds to certain ontological relations. Synonyms are substitutable in both directions. For example, *worker* and *employee* entail each other's meanings, as in the context "high salaried *worker/employee*". Hyponyms typically entail their hypernyms. For example, *dog* entails *animal*, as in "I have a *dog*" which entails "I have an *animal*" (but not vice versa). In some cases part-whole and member-set relations satisfy the meaning-entailing substitutability criterion. For example, a discussion of *division* entails in many contexts the meaning of *company*. Similarly, the plural form of *employee(s)* often entails the meaning of *staff*. On the other hand, non-synonymous words that share a common hypernym (co-hyponyms) like *company* and *government*, or *country* and *city*, are not substitutable since they always refer to different meanings (such as different entities).

Our test set included a sample of 30 randomly selected nouns whose corpus frequency is above 500. For each noun we computed the top 40 most similar words by both similarity measures, yielding a total set of about 1600 (different) suggested word similarity pairs. Two independent assessors were assigned, each judging half of the test set (800 pairs). The output pairs from both methods were mixed so the assessor could not relate a pair with the method that suggested it.

#Words	Judge 1 (%)	Judge 2 (%)	Total (%)
Top 10	63.4 / 54.1	62.6 / 53.4	63.0 / 53.7
Top 20	57.0 / 48.3	56.4 / 45.8	56.8 / 47.0
Top 30	55.3 / 45.1	53.3 / 43.4	54.2 / 44.2
Top 40	53.5 / 44.6	51.6 / 42.0	52.6 / 43.3

Table 4: Precision values for Top-N similar words by the *RFF* / Lin98 methods.

5.2 Similarity Results

The evaluation results are displayed in Table 4. As can be seen *RFF* outperformed Lin98 by 9-10 percentage points of precision at all Top-N levels, by both judges. Overall, *RFF* extracted 111 (21%) more correct similarity pairs than Lin98. The overall relative recall[3] of *RFF* is quite high (89%), exceeding Lin98 by 16% (73%). These figures indicate that our method covers most of the correct similarities found by Lin98, while identifying many additional correct pairs.

We note that the obtained precision values for both judges are very close at all table rows. To further assess human agreement level for this task the first author of this paper judged two samples of 100 word pairs each, which were selected randomly from the two test sets of the original judges. The proportions of matching decisions between the author's judgments and the original ones were 91.3% (with Judge 1) and 88.9% (with Judge 2). The corresponding Kappa values are 0.83 ("very good agreement") and 0.75 ("good agreement").

As for feature reduction, vector sizes were reduced on average to about one third of their original size in the Lin98 method (recall that standard feature reduction, tuned for the corpus, was already applied to the Lin98 vectors).

[2] Known as Reuters Corpus, Volume 1, English Language, 1996-08-20 to 1997-08-19.

[3] Relative recall shows the percentage of correct word similarities found by each method relative to the joint set of similarities that were extracted by both methods.

Feature	Weight
Industry, gen ↓	1.21
Airport, gen ↓	1.16
Neighboring, mod ↑	1.06
Law, gen ↓	1.04
Economy, gen ↓	1.02
Population, gen ↓	1.02
City, gen ↓	0.93
Impoverished, mod ↑	0.92
Governor, pcomp ↑	0.92
Parliament, gen ↓	0.91

Table 5: Top-10 features of *country* by *RFF*.

5.3 Empirical Observations for RFF

We now demonstrate the typical behavior of *RFF* relative to the observations and motivations of Section 3 (through the same example).

Table 5 shows the top-10 features of *country*. We observe (subjectively) that the list now contains quite indicative and reliable features, where too specific (anecdotal) and too general features were demoted (compare with Table 2).

More objectively, Table 6 shows that most of the top-10 common features for *country-state* are now ranked highly for both words. On the other hand, there are only two common features for the incorrect pair *country-economy*, both with quite low ranks (compare with Table 3). Overall, given the set of all the correct (judged as substitutable) word similarities produced by both methods, the average *top joint feature rank* of the top-10 common features by *RFF* is 21, satisfying the desired behavior which was suggested in Section 3. The same figure is much larger for the Lin98 vectors, which have an average top joint feature rank of 105.

Consequently, Table 7 shows a substantial improvement in the similarity list for *country*, where most incorrect words, like *economy* and *company*, disappeared. Instead, additional correct similarities, like *kingdom* and *land*, were promoted (compare with Table 1). Some semantically related but non-substitutable words, like "world" and "city", still remain in the list, but somewhat demoted. In this case all errors correspond to quite close semantic relatedness, being geographic concepts.

The remaining errors are mostly of the first type discussed in Section 3 – pairs of words that are

Country-State	Ranks		Country-Economy	Ranks	
Neighboring	3	1	Developed	50	100
Industry	1	11	Liberalization	100	79
Impoverished	8	8			
Governor	10	9			
Population	6	16			
City	17	18			
Economy	5	15			
Parliament	10	22			
Citizen	14	25			
Law	4	33			

Table 6: *RFF* weighting: Top-10 common features for *country-state* and *country-economy* along with their corresponding ranks in the two (sorted) feature vectors.

ontologically or thematically related but are not substitutable. Typical examples are co-hyponyms (*country - city*) or agent-patient and agent-action pairs (*industry – product, worker – job*). Usually, such word pairs also have highly ranked common features since they naturally appear with similar characteristic features. It may therefore be difficult to filter out such non-substitutable similarities solely by the standard distributional similarity scheme, suggesting that additional mechanisms are required.

6 Conclusions and Future Work

This paper proposed the following contributions:

1. Considering *meaning entailing substitutability* as a target goal and evaluation criterion for word similarity. This criterion is useful for many semantic-oriented NLP applications, and can be evaluated directly by human subjects.
2. A thorough empirical error analysis of state of the art performance was conducted. The main observation was deficient quality of the feature vectors which reduces the quality of similarity measures.
3. Inspired by the qualitative observations we identified a new qualitative condition for feature vector evaluation – *top joint feature rank*. Thus, feature vector quality can be measured independently of the final similarity output.
4. Finally, we presented a novel feature weighting function, *relative feature focus*. This measure was designed based on error analysis insights and im-

nation	1	territory	6	african_country	11	*district	16
state	2	*neighbor	7	province	12	european_country	17
island	3	colony	8	*city	13	zone	18
region	4	*port	9	*town	14	land	19
area	5	republic	10	kingdom	15	place	20

Table 7: Top-20 most similar words for *country* and their ranks in the similarity list by the *RFF*-based measure. Incorrect similarities (non-substitutable) are marked with '*'.

proves performance over all the above criteria.

We intend to further investigate the contribution of our measure to word sense disambiguation and to evaluate its performance for clustering methods.

Error analysis suggests that it might be difficult to improve similarity output further within the common distributional similarity schemes. We need to seek additional criteria and data types, such as identifying evidence for non-similarity, or analyzing more carefully disjoint features.

Further research is suggested to extend the learning framework towards richer notions of ontology generation. We would like to distinguish between different ontological relationships that correspond to the substitutability criterion, such as identifying the entailment direction, which was ignored till now. Towards these goals we plan to investigate combining unsupervised distributional similarity with supervised methods for learning ontological relationships, and with paraphrase acquisition methods.

References

Barzilay, Regina and Kathleen McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In Proc. of ACL/EACL, 2001.

Church, Kenneth W. and Hanks Patrick. 1990. Word association norms, mutual information, and Lexicography. *Computational Linguistics*, 16(1), pp. 22–29.

Dagan, Ido. 2000. Contextual Word Similarity, in Rob Dale, Hermann Moisl and Harold Somers (Eds.), *Handbook of Natural Language Processing*, Marcel Dekker Inc, 2000, Chapter 19, pp. 459-476.

Dagan, Ido and Oren Glickman. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. PASCAL Workshop on Text Understanding and Mining.

Dagan, Ido, Lillian Lee and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 1999, Vol. 34(1-3), special issue on Natural Language Learning, pp. 43-69.

Grefenstette, Gregory. 1994. Exploration in Automatic Thesaurus Discovery. *Kluwer Academic Publishers.*

Harris, Zelig S. 1968. *Mathematical structures of language.* Wiley, 1968.

Hindle, D. 1990. Noun classification from predicate-argument structures. In Proc. of ACL, pp. 268–275.

Lee, Lillian. 1997. Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis, Harvard University, Cambridge, MA.

Lin, Dekang. 1993. Principle-Based Parsing without Overgeneration. In Proc. of ACL-93, pages 112-120, Columbus, Ohio, 1993.

Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. In Proc. of COLING–ACL98, Montreal, Canada, August, 1998.

Lin, Dekang and Patrick Pantel. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7(4), pp. 343-360, 2001.

Pereira, Fernando, Tishby Naftali, and Lee Lillian. 1993. Distributional clustering of English words. In Proc. of ACL-93, pp. 183–190.

Ruge, Gerda. 1992. Experiments on linguistically-based term associations. *Information Processing & Management*, 28(3), pp. 317–332.

Weeds, Julie and David Weir. 2003. A General Framework for Distributional Similarity. In Proc. of EMNLP-03. Spain.