Modeling Semantic Coherence from Corpus Data: The Fact and the Frequency of a Co-occurrence

Viktor Pekar
Bashkir State University

1. Introduction

One puzzling property of natural language is its so-called infinite productivity, in other words, the ability of a linguistic item to combine with an infinite number of other linguistic items in a meaningful way. From a computational perspective, the task of modeling this phenomenon amounts to estimating the plausibility of the combination of any two or more given linguistic items. Satisfactory solutions to this task are of relevance for various areas in natural language processing: speech recognition, syntactic attachment resolution, word sense disambiguation, information retrieval, etc. Corpus-based models are particularly capable of accomplishing this task: using no other data than corpora they can automatically form a representation of a given item, generalize it to a linguistic (morphological, syntactic, semantic) category and later handle the item using available knowledge about the category. The present paper addresses problems of estimating semantic coherence between words in a phrase.

Out of the corpus-based models, the so-called similarity-based models display a better ability to account for finer semantic distinctions, rather than thesaurus-based models. The latter makes use of predefined semantic classes of words, such as synonym groups in WordNet (e.g. Li and Abe 1998). This approach assumes that all words within a class have similar co-occurrence features and that the classes have clear-cut boundaries. This assumption is reasonable when only crude semantic distinctions are sufficient, for example, in the task of syntactic attachment resolution. However, if one is concerned with finer semantic distinctions, this approach is not adequate, as a word, on the one hand, has semantic dissimilarities with other words within its class and, on the other hand, can be semantically close to words of other classes (Schütze 1997). The similarity-based approach does not use any precompiled semantic classes; instead, they are computed from corpus data and therefore they display a gradient character[1]. The gradient character of the thus derived category is cognitively plausible: it reflects the fact that people display different degrees of certainty about the appropriateness of a combination of a word with different members of the same category.

The general design of a similarity-based model of a word-co-occurrence can be described as follows. A word w^j is represented in terms of words $w^i_{1...n}$ with which w^j co-occurs. For deriving a representation of w^j, the application searches through the corpus for its occurrences and counts words occurring in the context of w^j. To characterize the context of w^j, two strategies have been tried: the window-based and the syntactic. In the first case the context is marked out by imposing a window of a certain size around w^j (e.g., Gale, Church, and Yarowsky (1992) used a thousand-word window) The other strategy is to limit context to words appearing in a certain syntactic relation to w^j, such as direct objects of a verb (Grefenstette 1996; Pereira, Tishby, and Lee 1993).

After the data are gathered, the semantics of w^j is represented in terms of a vector in an n-dimensional

[1] The classes are formed on the bases of distributional data, which do not always correctly reflect the semantics of words (Charniak 1993:144-145). Nonetheless, they can be effectively used to model semantic coherence between words. The expressions "semantic class" and "distributional class" will therefore be used interchangeably in this paper.

Coyote Papers 12, 1-8
Language in Cognitive Science

space, where n is the number of words co-occurring with w^j and components of the vector are probabilities of the co-occurrences established from their observed frequencies:

$C(w^j) = \langle P(w^j|w^i_1), P(w^j|w^i_2), P(w^j|w^i_3), \ldots P(w^j|w^i_n)\rangle$.

The words $w^i_{1\ldots n}$ in the vector have different relevance for representing w^j. For example, if one considers function words in the context, such as particles and conjunctions, they are likely to have a high frequency of co-occurrence, but they do not tell much about the semantics of a specific word. To represent w^j more accurately, one can calculate relevance of each word and accordingly give weight to each component in the vector (for example, Schütze (1992) used Canonical Discriminant Analysis technique for this purpose). The obtained vector can be interpreted geometrically: it points in a unique direction thus reflecting unique semantics of a word.

Given that vectors of individual words are in the same vector space, the meaning of the words can be compared considering distance between the vectors. There are a variety of methods for measuring semantic similarity between two words by considering their distributional representation. Similarity can be measured by (1) Euclidean distance between the vectors of the words, (2) the cosine between the vectors (Schütze 1997); (3) Kullback-Leibler divergence (Pereira, Tishby, and Lee 1993); (4) Jensen-Shannon divergence (Dagan, Lee, and Pereira 1999); etc.

Using the distributional representation and a similarity measuring technique, it is possible to estimate the likelihood of the combination of any two given words w_1 and w_2, even if the corpus does not contain instances of this combination. It can be done through supplying missing distributional data about w_1 by data about words w_1' that are semantically close to w_1. There are two main approaches to establish how probable an unseen combination is: the word-based, or nearest neighbors method (Dagan, Marcus, and Markovitch 1995; Dagan, Lee, and Pereira 1999) and the class-based method (Pereira, Tishby, and Lee 1993).

The nearest neighbors method consists in determining words w_1' that are most similar to w_1 out of all the words that were seen to combine with w_2. After these words are determined, the likelihood of w_2 given w_1 is established from a sum of all $P(w_2|w_1')$ for each w_1', each of the probabilities being weighted by the similarity measure between w_1 and w_1'. To ensure that contribution of only closest words is non-negligible, the weight is presented as an exponential function with the similarity measure multiplied by a free parameter as an exponent.

The class-based method is different from the nearest neighbors method in that it determines words w_1' not on the basis of their similarity to w_1, but through clustering the vectors of all words appearing in a certain syntactic relation to w_2. A centroid of each cluster is calculated by averaging distributional representations of the words making up the cluster. The words w_1' appear in that cluster, the centroid of which is the closest to the vector of w_1. The probability that w_1 co-occurs with w_2 is derived from the degree of closeness between the vector of w_1 and the centroid which is closest to it. Thus the nearest neighbors method can be said to form semantic classes in a context-sensitive manner, while the class-based method makes use of permanently existing semantic classes of a language.

The main problem that both models have to deal with is data sparseness. It consists in the fact that in a given corpus there may be no co-occurrences of w^j with some of w^i_n that are expected to characterize w^j. As a result, two words that are normally perceived to be similar may have vectors pointing in very different directions. To counter this problem, one has to use smoothing techniques, such as the back-off method (Dagan, Lee, and Pereira 1999), which requires significant computational expenses.

Another problem is that the vector space in these models normally comes out extremely large, because one has to ensure that all the words that need to be compared are represented along the same dimensions. The vector space is computationally very demanding even if one limits context of w^j to words standing in a syntactic relation to it, rather than using a window on the order of a thousand words. Therefore, application of dimensionality reduction techniques, such as singular value decomposition (e.g., Schütze 1997), is needed to make computation less complex. Besides, as Dagan, Lee, and Pereira (1999) noted, dimensionality reduction might result in better generalization ability of the model, although a reliable experimental proof for this has not been obtained.

The present paper proposes a format for representing vectors of words that might help, on the one hand, to minimize negative effects of data sparseness and, on the other hand, to reduce the vector space. Thematic roles of

verbs and semantics of nouns will be represented in terms of this format. The model's ability to predict semantic cohesion between the verbs and the nouns will be checked against human subjects' judgements about the appropriateness of filling thematic roles of the verbs with different nouns. The rest of the paper is organized as follows. Section 2 describes the proposed representation. Section 3 presents the design of the model. Section 4 deals with its experimental evaluation and discussion. Section 5 outlines future directions of the research.

2. The Fact of Co-occurrence versus the Frequency of Co-occurrence

Both the similarity-based and the class-based methods manipulate corpus data presented in terms of probabilities, whereby weight is given to number of co-occurrences rather than to the fact itself that the words co-occur (or do not co-occur). As a result, zero probability, usually signifying that the words cannot co-occur, has the same relevance as any other value. For example, the difference between a one-count bigram and a zero-count bigram is treated the same as the difference between a one-count bigram and a two-count bigram. The models are thus poor at distinguishing between the possibility and impossibility of a combination. On the other hand, reliance on probability (and ultimately on frequency) makes two words different, if they have positive, but differing probability values. It is clear that once one knows that two words co-occur regularly, the exact number of their co-occurrences does not make much difference. To avoid great accidental differences in frequencies, the models need to employ larger corpora and use smoothing techniques.

In the project described here, I am trying to check the hypothesis that the semantics of a word can be adequately reflected by information about the possibility/impossibility of its co-occurrence with other words without using frequency information. Components of the vector of a word w^j in this model will thus be only specific words $w^i_{1..n}$ which are observed to co-occur with it. Generalization over several such vectors amounts to establishing which components are common to these vectors while ignoring differing components. The resulting generalized representation will consist of only those components that are common to all the compared vectors and thus it will be more abstract than the vectors over which it was generalized. This representation can be expected to have better generalizing ability, i.e. ability to be adequately introduced into novel contexts, than the averaged representation used in the class-based model. In the class-based model, averaging over several vectors consists in deriving a centroid, i.e. in summing probabilities of corresponding components and normalizing the result by the number of the vectors. This procedure presupposes that zero probabilities are summed with other values so that the averaged representation retains the quality of being as specific as each of the separate vectors. The averaged representation, as opposed to the abstract one, misses the fact that some distributional features are important in representing a linguistic item, while some are redundant. It can be expected that trying to detect these redundant features in a novel context causes incorrect predictions about the usage of the word.

It might be argued that the proposed representation is inadequate, because words w^i infrequently co-occurring with w^j, such as personal nouns, have the same relevance as words w^i with a high frequency of co-occurrence with w^j. However, low-frequency co-occurrences may in fact make an important contribution to the representation of words. Dagan, Lee and Pereira (1999) found that deleting co-occurrences that are present only once in the corpus from training data results in a significant change in the similarity measures between words. They attributed this to the fact that out of all the verbs in their corpus, 65% were those that co-occurred with 10 or fewer nouns. Indeed, removing some of these nouns will significantly change the representation of these verbs.

To sum up, the proposed method can have the following advantages over the frequency-based representations. First, it allows for a more acceptable degree of data sparseness, since differences in counts between co-occurrences may be ignored. Second, the vector space is represented in a more compact way through omitting frequency information and removing components with zero counts. Third, the generalized representation may have more generalizing power through losing some specificity.

The proposed representation, however, can be plausible only on the condition that the context of w^j is defined not by a window of words, but by words that are syntactically related to it. As Schütze (1992) observed, when one uses a window of about a thousand words, the resulting representation for a word is so big that only

about 10% in the typical 4000-by-4000 matrix are zeros. Hence, for the window-based approach, observing just co-occurring words while neglecting their frequency cannot work. Therefore, the context of a word is chosen to be represented in terms of words, which are syntactically related to it.

In the model described in this paper, the contexts of verbs are represented in terms of their observed syntactic arguments and contexts of nouns–in terms of verbs they combine with. In building vectors for thematic roles of verbs, however, syntactic arguments are not differentiated, since one and the same thematic role can have different syntactic realizations. To derive a representation of a thematic role, all arguments of the verb are first collapsed together and then they are clustered according to their semantic similarity, the number of the clusters being the same as the number of syntactic arguments of the verb[2]. After that individual representations of nouns forming a cluster are generalized.

Semantics of nouns is likewise represented in terms of unique words, which co-occur with the nouns, irrespective of the frequency of the co-occurrences. Later, to determine the likelihood that a given noun fills a given thematic role of a verb, representation of the noun is compared to the generalized representation of a thematic role.

3. Training data and algorithm

For a preliminary estimation of the viability of the model, it was trained on a manually prepared corpus. Its format simulated data that can be extracted from a syntactically annotated corpus: each record in the database represented the argument structure of a verb and contained the verb and up to 3 arguments of the verb. The size of the database was 758 records, or about 2500 words. The training data contained 160 individual words (61 verbs and 99 nouns), i.e. each of the verbs was combined with no nouns other than the ones present in the database. The number of occurrences for each word ranged between 3 and 92.

Table 1. An example of records in the database.

VERB	ARG1	ARG2	ARG3
give	*John*	*Peter*	*book*
read	*John*	*book*	-
write	*John*	*Mary*	*letter*

To form a representation of a particular noun the application made use of the following algorithm:

1) Records containing the noun are extracted.
2) A vector of the noun is represented in terms of unique words it was seen to combine with.

To derive a generalized representation of thematic roles of a verb, the following subtasks are carried out:

1) Records containing the same number of arguments of the verb are extracted.
2) Based on similarity of their vectors, the nouns are arranged into clusters, the number of clusters being the same as the number of arguments in the records.
3) Vectors of nouns forming a cluster are compared to filter out differing features; features that are common to all nouns in the cluster constitute a generalized representation of a thematic role.

In measuring similarity between two vectors, the metrics used was a binary Tanimoto measure (Charniak 1993:142), which is the ratio between the number of attributes shared by w_1 and w_1' and the number of unique

[2] Neglecting syntactic details causes, of course, loss of some semantic information. In particular, it does not allow one to differentiate between two different arguments with very similar semantics (e.g. the subject and the object of *to kill* in *John killed Peter*.). Such cases require the use of additional techniques of mapping syntactic arguments to thematic roles and vice versa.

attributes possessed by w_1 and w_1' . Attributes in the denominator are those of the word that has the greater number of attributes. Thus, the value of the measure is always between 0 and 1. Table 2 gives examples of words in the corpus found similar to the verb *eat* using this similarity measure.

Table 2. Verbs found similar to *eat*.

Word	Similarity measure
like	0.24166
buy	0.23497
sell	0.18037
hate	0.14333
gather	0.12251
bite	0.12152
drink	0.11944
grow	0.10139

4. Model evaluation

A representation of a thematic role can be considered adequate, if it satisfies the following two conditions. It must be (1) general enough to allow novel arguments to fit into it and (2) specific enough to reflect individual semantics of the verb and disallow inappropriate combinations. To test the adequacy of the obtained representations of thematic roles, two experiments were conducted that consisted in filling thematic roles with different nouns and obtaining human subjects' judgements about the appropriateness of the resulting combinations. The likelihood of a noun's filling a thematic role of a verb was estimated by measuring similarity between the vector of the noun and the generalized representation of the thematic role (the assumption is that the more attributes the noun has in common with the thematic role, the greater likelihood that the noun can fill it)[3]. Thus, validating the adequacy of the representations amounts to demonstrating a correlation between the human evaluations and the probabilities calculated by the model. Greater probabilities calculated by the model would be expected to be related to higher human evaluations, and lower probabilities would be expected to be related to lower human evaluations.

During either of the experiments, the application combined one verb and one noun so that the produced combinations were absent in the corpus. The noun was tried in each of the thematic roles of the verb[4]. The rest of the roles were filled by nouns, which combined with the verb in the corpus. The subjects evaluated the produced verb-argument structures in terms of three ratings: "good", "doubtful", and "bad". After that disparities between the three groups of estimated verb-argument structures were assessed in terms of the chi-square criterion and the Wilcoxon (Mann-Whitney) criterion. In calculating the chi-square criterion, the entire space of possible values of EM was divided into equal ranges of 0.05, which were used as classes. The number of the verb-argument structures in a range was used as the frequency of a class (cf., Figures 1 and 2).

Experiment 1 consisted of combining a specific verb with each of the 99 nouns, that is, all of the possible combinations of the verb with the nouns were included in the test sample. The training data were made up of the verbs *break*, *cut*, and *put*, which appeared 7, 18, and 30 times, respectively, in the corpus. The test data of the experiment were 99 verb-argument structures (VA structures) generated by the model, i.e. the combinations of a verb with each of the 99 nouns. Table 3 details data from Experiment 1: for each type of the subjects' ratings, the number of generated verb-argument structures and the means of their EM are given. The numbers in parentheses after the verbs specify the number of occurrences of the verbs in the corpus.

---

[3] In the following, to denote the similarity measure between the candidate noun and the thematic role, which is used as a means to estimate the likelihood of the combination, the expression "estimation measure" (EM) will be used.

[4] Here the same measuring technique was used as in comparing two individual words, i.e. Tanimoto measure. However, the attributes in the denominator were always those of the thematic role.

Table 3. Test data for Experiment 1.

Evaluation	*break* (7)		*cut* (18)		*put* (30)	
	number	means	number	means	number	means
good	28	0.19785	34	0.23617	48	0.2656
doubtful	24	0.14166	9	0.1425	9	0.1712
bad	47	0.11382	56	0.10875	42	0.1

Table 4 specifies the parameters of statistical reliability of the disparities between the three groups of EM of generated VA structures with the verb *put*.

Table 4. Disparities between subjects' ratings of VA structures with the verb *put*.

Compared groups	χ^2		Mann-Whitney			
	χ^2	α	W_{lowest}	$W_{observed}$	$W_{highest}$	α
"good" vs. "doubtful"	207.6499*	<0.01	1262	1527	1522	0.005
"good" vs."bad"	1566.853*	<0.01	1763	3116.5	2605	0.001
"doubtful" vs."bad"	233.671*	<0.01	103	365	365	0.002

* $v_2 = 1$.

Figure 1 presents a histogram reflecting distribution of generated VA structures for each of the subjects' ratings in Experiment 1. The x-axis specifies the ranges of EM. The y-axis specifies the frequencies of appearance of the VA structures within these ranges as percentages of the total number of the generated VA structures.

During Experiment 2, the application chose at random one verb and one noun out of all the words of the corpus, i.e. the possible combinations amounted to 99 nouns x 61 verbs x at least two argument positions for each verb = 12078. The test data consisted of 631 VA structures, in which the verb and the noun were chosen at random. The training data were all of 758 sentences of the corpus. Table 5 summarizes the test data. For each type of human rating, it gives means of EM for VA structures with this rating.

Table 5. Test data for Experiment 2.

Ratings	Number of VA structures	Means of estimation measures
good	92	0.22782
doubtful	60	0.12016
bad	479	0.7901

Table 6 specifies the parameters of statistical reliability of the disparities between the three groups of the generated VA structures in Experiment 2.

Table 6. Disparities between subjects' ratings of VA structures in Experiment 2.

Compared groups	χ^2		Mann-Whitney			
	χ^2 value	α	W_{lowest}	$W_{observed}$	$W_{highest}$	α
"good" vs. "doubtful"	89.45384*	<0.01	6135	8727	7941	0.001
"good" vs."bad"	168.2707*	<0.01	21383	44668	31241	0.001
"doubtful" vs."bad"	15.94737**	<0.01	14334	18421	18066	0.1

* $v_2 = 5$; ** $v_2 = 3$.

Figure 2 presents a histogram reflecting the distribution of the generated VA structures for each of the subjects' ratings in Experiment 2. The x-axis specifies the ranges of the estimation measure, the y-axis specifies the frequencies of appearance of the VA structures within these ranges as a percentage of the total number of the generated VA structures.

The statistical evaluation of the data of the both experiments indicate that the disparities between the three groups of VA structures corresponding to the three types of subjects' ratings are statistically reliable, i.e. there is a correlation between the value of EM calculated by the model and the ratings given by the subjects. The model is able to make a clear distinction between combinations evaluated as "good" and those evaluated as "bad" by human subjects, as well as between "good" and "doubtful" combinations. The distinction between "doubtful" and "bad" is less clear.

As is seen from the data in Figure 2, combinations evaluated as "bad" very seldom have EM above 0.15 (in 4% of the cases). 56% of combinations evaluated as "good" have EM above 0.15. The fact that 44% are below this level should be accounted for by sparseness of the data: the corpus lacked some argument structure patterns that are quite plausible. For example, there were no argument structures in the corpus describing humans possessing certain types of artifacts. It can be expected that after enlarging the corpus more VA structures evaluated as "good" will have greater EM.

The clear distinction made by the model between "good" and "bad" human judgements as well as the high ratio between test and training data in both experiments (for *break*: 14.143 (99/7), for *cut*: 5.5 (99/18), for *put*: 3.09 (99/32), for the second experiment: 0.832 (631/758)) indicate that the calculated thematic roles of verbs indeed satisfy the criteria of being both general enough to let novel nouns fill them and specific enough to disallow inappropriate combinations.

6. Conclusion and future directions

In this work, a format for distributionally representing meanings of individual words and a method of generalization over such representations were proposed. The format gives weight to the fact that words co-occur rather than to the frequency of their co-occurrences. This representation was hypothesized to have a number of advantages over frequency-based representations: the ability to be trained on data with a greater degree of sparseness, reduction of vector space, and improved generalizing power. The model was assessed by comparing its evaluations of semantic coherence between a verb and a noun with human subjects' ratings of these combinations. The results indicate that there is a correlation between the estimation measures of the combinations calculated by the model and the human subjects' judgements, thus showing that, in majority of cases, the model is able to predict semantic well-formedness of novel word combinations, at least using the present corpus. Further studies, however, are required to establish that the format indeed has the hypothesized advantages.

The model will be trained on a bigger amount of data from a larger corpus and later evaluated using the same experimental procedure. The proposed format of representation can be studied in several directions. The use of a larger corpus will allow studying the effect of the low-frequency co-occurrences on the proposed model. In

particular, it can be done by removing low-frequency co-occurrences from the training set and later comparing the model' s performance on the two kinds of training data–with and without the low-frequency co-occurrences. Furthermore, one can study the effect of low-frequency co-occurrences with different number of counts to empirically establish the character of dependence between the number of counts of context words and the model' s performance.

To obtain further evidence for advantages and disadvantages of the proposed format, it will be compared to the frequency-based representations after the both types have been trained on the same corpus. In particular, the effect of data sparseness on the two models can be studied and the computational cost of the two models can be compared. A hybrid representation may also be developed, which would include information about the possibility/impossibility of a co-occurrence, while also using frequency information as weights for particular features in a vector. The value of a feature of a vector can be presented as an exponential function so that differences between high-frequency co-occurrences are made negligible, whereas differences between low-frequency ones are made prominent; the zero frequency co-occurrences, however, are kept distinct from non-zero ones. To establish the limit where frequency data start to lose relevance, one can make use of an experimentally tunable parameter multiplying the exponent in the weight function. The hybrid representation will need to be compared to the frequency-based and the fact-based representations.

References

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.

Dagan, I., Lee, L., Pereira, F. (1999). Similarity-based models of word co-occurrence probabilities. *Machine Learning* 34(1-3): 43-69.

Dagan, I., Marcus, S., Markovitch, S. (1995). Contextual word similarity and estimation from sparse data. *Computer Speech and Language*, 9: 123-152.

Gale, W. A., Church, K.W., Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*.

Grefenstette, G. (1996). Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In B. Boguarev and J. Pustejovsky (eds.). *Corpus Processing for Lexical Acquisition.* Cambridge, MA: MIT Press.

Li, H., Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics* 24(2): 217-226.

Pereira, F., Tishby, N., Lee., L. (1993). Distributional clustering of English words. In *Proceedings of the Thirty-first Annual Meeting of the ACL*, 183-190.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of Supercomputing 92*.

Schütze, H. (1997). *Ambiguity resolution in language learning: Computational and Cognitive Models*. Stanford, CA: CSLI Publications.

BSPU
Okt.Revolutsii,
3a, Ufa 450000, Russia.

vpekar@ufanet.ru