19. CONTEXTUAL WORD SIMILARITY

Ido Dagan

Mathematics and Computer ScienceDepartment

Bar Ilan University

Ramat Gan 52900, Israel

dagan@cs.biu.ac.il

19.1. Introduction

Identifying different types of similarities between words has been an important goal in Natural Language Processing (NLP). This chapter describes the basic statistical approach for computing the degree of similarity between words. In this approach a word is represented by a **word co-occurrence vector** in which each entry corresponds to another word in the lexicon. The value of an entry specifies the frequency of joint occurrence of the two words in the corpus, that is, the frequency in which they co-occur within some particular relationships in the text. The degree of similarity between a pair of words is then computed by some similarity or distance measure that is applied to the corresponding pairs of vectors.

This chapter describes in detail different types of lexical relationships that can be used for constructing word co-occurrence vectors. It then defines a schematic form for vector-based similarity measures and describes concrete measures that correspond to the general form. We also give examples for word similarities identified by corpus-based measures, along with "explanations" of the major data elements that entailed the observed similarity.

This chapter describes in detail the basic vector-based approach for computing word similarity. Similarities may be computed between different lexical units, such as word strings, word lemmas, and multi-word terms or phrases. We shall use the term "word" to denote a lexical unit but the discussion applies to other units as well.

19.1.1. Applications of word similarity

The concept of word similarity was traditionally captured within thesauri. A thesaurus is a lexicographic resource that specifies semantic relationships between words, listing for each word related words such as synonyms, hyponyms and hypernyms. Thesauri have been used to assist writers in selecting appropriate words and terms and in enriching the vocabulary of a text. To this end, modern word processors provide a thesaurus as a built in tool.

The area of information retrieval (IR) has provided a new application for word similarity in the framework of **query expansion**. Good free-text retrieval queries are difficult to formulate since the same concept may be denoted in the text by different words and terms. Query expansion is a technique in which a query is expanded with terms that are related to the original terms that were given by the user, in order to improve the quality of the query. Various query expansion methods have been implemented, both by researchers and in commercial systems, that rely on manually crafted thesauri or on statistical measures for word similarity (Frakes, 1992).

Word similarity may also be useful for disambiguation and language modeling in the area of NLP and speech processing. Many disambiguation methods and language models rely on word co-occurrence statistics that are used to estimate the likelihood

of alternative interpretations of a natural language utterance (in speech or text). Due to data sparseness, though, the likelihood of many word co-occurrences cannot be estimated reliably from a corpus, in which case statistics about similar words may be helpful.

Consider for example the utterances in (1), which may be confused by a speech recognizer.

(1) a. The bear ran away.

b. The pear ran away.

A typical language model may prefer the first utterance if the word co-occurrence *bear ran* was encountered in a training corpus while the alternative co-occurrence *pear ran* was not. However, due to data sparseness it is quite likely that neither of the two alternative interpretations was encountered in the training corpus. In such cases information about word similarity may be helpful. Knowing that *bear* is similar to other animals may help us collect statistics to support the hypothesis that animal names can precede the verb *ran*. On the other hand, the names of other fruits, which are known to be similar to the word *pear*, are not likely to precede this verb in any training corpus. This type of reasoning was attempted in various disambiguation methods, where the source of word similarity was either statistical (Grishman et al., 1986; Schutze, 1992, 1993; Essen and Steinbiss, 1992; Grishman and Sterling, 1993; Dagan et al., 1993, 1995; Karov and Edelman, 1996; Lin, 1997) or a manually crafted thesaurus (Resnik, 1992, 1995; Jiang and Conrath, 1997).

It should be noted that while all the applications mentioned above are based on some notion of "word similarity" the appropriate type of similarity relationship might vary. A thesaurus intended for writing assistance should identify words that resemble each other in their meaning, like *aircraft* and *airplane*, which may be substituted for each other. For query expansion, on the other hand, it is also useful to identify contextually related words, like *aircraft* and *airline*, which may both appear in relevant target documents. Finally, co-occurrence-based disambiguation methods would benefit from identifying words that have similar co-occurrence patterns. These might be words that resemble each other in their meaning, but may also have opposite meanings, like *increase* and *decrease*.

19.1.2. The corpus-based approach

Traditionally, thesauri have been constructed manually by lexicographers. Like most lexicographic tasks, manual thesaurus construction is very tedious and time-consuming. As a consequence, comprehensive thesauri are available only for some languages, and for the vocabularies of very few professional domains, such as the medical domain. Furthermore, while vocabularies and their usage patterns change rapidly, the process of updating and distributing lexicographic resources is slow and lags behind the evolution of language. These problems may be remedied, at least to some extent, by automatic and semi-automatic thesaurus construction procedures that are based on corpus statistics.

The common corpus-based approach for computing word similarity is based on representing a word (or term) by the set of its **word co-occurrence** statistics. It relies on the assumption that the meaning of words is related to their patterns of co-occurrence with other words in the text. This assumption was proposed in early linguistic work, as expressed in Harris' **distributional hypothesis**: "… the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities." (Harris, 1968, p. 12). The famous statement "You shall know a word by the company it keeps!" (Firth, 1957, p. 11) is another expression of this assumption.

Given the distributional hypothesis, we can expect that words that resemble each other in their meaning will have similar co-occurrence patterns with other words. For example, both nouns *senate* and *committee* co-occur frequently with verbs like *vote, reject, approve, pass* and *decide*. To capture this similarity, each word is represented by a **word co-occurrence vector**, which represents the statistics of its co-occurrence with all other words in the lexicon. The similarity of two words is then computed by applying some vector similarity measure to the two corresponding co-occurrence vectors. Section 19.2 describes how different types of word co-occurrence vectors can be constructed by considering different types of co-occurrence relations (Section 19.2.3 relates to the alternative representation of **document occurrence vector**, which is commonly used in information retrieval).

Section 19.3 defines a schematic form for vector-based similarity measures while Section 19.4 describes concrete measures and discusses their correspondence to the schematic form. Section 19.5 lists examples for word similarities identified by corpus-based measures, along with "explanations" of common co-occurrences patterns that entailed the observed similarity.

19.2. Co-occurrence relations and co-occurrence vectors

In the corpus-based framework a word is represented by data about its joint co-occurrence with other words in the corpus. To construct representations, we should first decide what counts as a co-occurrence of two words and specify informative types of relationships between adjacent occurrences of words. Different types of co-occurrence relationships have been examined in the literature, for computing word similarity as well as for other applications. These relationships may be classified into two general types: grammatical relations, which refer to the co-occurrence of words within specified syntactic relations, and non-grammatical relations, which refer to the co-occurrence of words within a certain distance (window) in the text. As will be discussed below, the types of relations used in a particular word similarity system will affect the types of similarity that will be identified.

19.2.1. Grammatical relations

Lexical co-occurrence within syntactic relations, such as subject–verb, verb–object and adjective–noun, provide an informative representation for linguistic information. Statistical data on co-occurrence within syntactic relations can be viewed as a statistical alternative to traditional notions of selectional constraints and semantic preferences. Accordingly, this type of data was used successfully for various broad coverage disambiguation tasks.

The set of co-occurrences of a word within syntactic relations provides a strong reflection of its semantic properties. It is usually the meaning of the word which restricts the identity of other words that can co-occur with it within specific syntactic relationships. For example, *edible* items can be the direct object of verbs like *eat, cook,* and *serve,* but not of other verbs such as *drive.* When used for assessing word similarity, grammatical co-occurrence relationships are likely to reveal similarities between words that share semantic properties.

In order to extract syntactically based lexical relations it is necessary to have a syntactic parser. Although the accuracy of currently available parsers is limited, it was shown that limited accuracy is sufficient for acquiring reliable statistical data, with a rather limited amount of noise that can be tolerated by the statistical similarity methods. Yet, the use of a robust parser may be considered as a practical disadvantage in some situations, since such parsers are not yet widely available and may not be sufficiently efficient or accurate for certain applications.

19.2.2. Non-grammatical relations

Non-grammatical co-occurrence relations refer to the joint occurrence of words within a certain distance (window) in the text. This broad definition captures several sub-types of co-occurrence relations such as n-grams, directional and non-directional co-occurrence within small windows, and co-occurrence within large windows or within a document.

19.2.2.1. N-grams

An n-gram is a sequence of *n* words that appear consecutively in the text. N-gram models are used extensively in language modeling for automatic speech recognition systems (see Jelinek et al. (1992) for a thorough presentation of this topic), as well as in other recognition and disambiguation tasks. In an n-gram model the probability of an occurrence of a word in a sentence is approximated by its probability of occurrence within a short sequence of *n* words. Typically sequences of two or three words (bigrams or trigrams) are used, and their probabilities are estimated from a large corpus. These probabilities are combined to

estimate the *a priori* probability of alternative acoustic interpretations of the utterance in order to select the most probable interpretation.

The information captured by n-grams is, to a large extent, only an indirect reflection of lexical, syntactic and semantic relationships in the language. This is because the production of consecutive sequences of words is a result of more complex linguistic structures. However, n-grams have been shown to have practical advantages for several reasons: it is easy to formulate probabilistic models for them, they are very easy to extract from a corpus, and, above all, they have proved to provide useful probability estimations for alternative readings of the input.

Word similarity methods that are based on bigram relationships were tried for addressing the data sparseness problem in n-gram language modeling (Essen & Steinbiss, 1992; Dagan et al., 1994). Word-similarities that are obtained by n-gram data may reflect a mixture of syntactic, semantic, and contextual similarities, as these are the types of relationships represented by n-grams. Such similarities are suitable for improving an n-gram language model, which, by itself, mixes these types of information.

19.2.2.2. Co-occurrence within a large window

A co-occurrence of words within a relatively large window in the text suggests that both words are related to the general topic discussed in the text. This hypothesis will usually hold for frequent co-occurrences, that is, for pairs of words that often co-occur in the same text. A special case for this type of relationship is co-occurrence within the entire document, which corresponds to a maximal window size.

Co-occurrence within large windows was used in the work of Gale et al. (1993) on word-sense disambiguation. In this work co-occurrence within a maximal distance of 50 words in each direction was considered. A window of this size captures context words that identify the topic of discourse. Word co-occurrence within a wide context was used also for language modeling in speech recognition in the work of Lau et al. (1993), where the occurrence of a word affects the probability of other words in the larger context. In the context of computing word similarity, co-occurrence within a large window may yield topical similarities between words that tend to appear in similar contexts.

19.2.2.3. Co-occurrence within a small window

Co-occurrence of words within a small window captures a mixture of grammatical relations and topical co-occurrences. Typically, only co-occurrence of content words is considered since these words carry most semantic information.

Smadja (1993) used co-occurrence within a small window as an approximation for identifying significant grammatical relations without using a parser. His proposal relies on an earlier observation that 98% of the occurrences of syntactic relations relate words that are separated by at most five words within a single sentence (Martin, 1983). Smadja used this fact to extract lexical collocations, and applied the extracted data to language generation and information retrieval. Dagan et al. (1993, 1995) use this type of data as a practical approximation for extracting syntactic relationships. To improve the quality of the approximation, the direction of co-occurrence is considered, distinguishing between co-occurrences with words that appear to the left or to the right of the given word. The extracted data is used to compute word similarities, which capture both semantic similarities, as when using grammatical relations, but also some topical similarities, as when using co-occurrence within a larger context.

Another variant of co-occurrence within a small window appears in the work of Brown et al. (1991). They use a part-of-speech tagger to identify relations such as “the first verb to the right” or “the first noun to the left”, and then use these relations for word-sense disambiguation in machine translation. This type of relationship provides a better approximation for syntactically motivated relations while relying only on a part of speech tagger, which is a simpler resource compared to syntactic parsers.

19.2.3. Word co-occurrence vectors

The first step in designing a word similarity system is to select the type of co-occurrence relations to be used, and then to choose a specific set of relations. For example, we may decide to represent a word by the set of its co-occurrences with other words within grammatical relations, as identified by a syntactic parser, and then select a specific set of relations to be used (subject–verb, verb–object, adjective–noun, etc.). Alternatively, we can use co-occurrences within a small window of three words, and distinguish between words that occur to the left or to the right of the given word. Notice that in both cases there is more than one type of relation to be used. For example, when representing the word *boy*, and considering its co-occurrences with the verb *see*, the representation should distinguish between cases in which *boy* is either the subject or the object of *see*.

In formal terms, a word u is represented by a set of **attributes** and their frequencies. An attribute, $att=<w,rel>$, denotes the co-occurrence of u with another word w within a specific relationship *rel*. For example, when representing the word *boy*, its occurrences as the subject and the object of the verb *see* will be denoted by the two attributes <*see,subj*> and <*see,obj*>. The word u is represented by a **word co-occurrence vector** that has an entry for each possible attribute of u. The value of each entry is the frequency in which u co-occurs with the attribute, denoted by *freq(u,att)*. The co-occurrence vector of u thus summarizes the statistics about the co-occurrences of u with other words, while representing the particular types of relationships that are distinguished by the system. In the special case in which only one type of relationship is used, such as the relationship "the word preceding u" (bigram), the relation *rel* can be omitted from the attribute, which will consist only of the word w. In this case, the length of the co-occurrence vector is equal to the number of words in the lexicon.

Co-occurrence vectors are typically sparse since a word typically co-occurs only with a rather small subset of the lexicon, and hence with a small set of attributes. Accordingly, the co-occurrence vector of u is represented in memory as a sparse vector which specifies only **active** attributes for u, that is, attributes which actually co-occur with it in the corpus. A sparse co-occurrence vector may also be considered as a set of attribute-value pairs, where the value of an attribute is its co-occurrence frequency with u (notice that the order of entries in the co-occurrence vector is not important as long as it is the same for all vectors). This chapter uses the notation of vectors rather than attribute–value pairs, which is the common notation in the literature of IR and statistical NLP.

The similarity and distance measures described in the following sections take as input two co-occurrence vectors, which represent two words, and measure the degree of similarity (or dissimilarity) between them. The vector representation is general and enables us to separate the two major variables in a word similarity system: the types of co-occurrence relations for the attributes of the vectors and the type of similarity measure that is used to compare pairs of vectors. These two variables are orthogonal to each other: any type of co-occurrence relation can be combined with any type of similarity measure. The set of co-occurrence relations that is chosen affects the nature of similarities that will be identified, such as semantic similarity versus topical similarity. The concrete similarity measure that is chosen affects the resulting similarity values between different word pairs. For example, the most similar word to u may be v_1 according to one similarity measure and v_2 according to another measure, where both values are based on the same vector representation.

The IR literature suggests another type of vector representation for computing word similarities (Frakes, 1992). In this framework a word u is represented by a **document occurrence vector**, which contains an entry for each document in the corpus. The value of an entry denotes the frequency of u in the corresponding document. Word similarity is then measured by the degree of similarity between such vectors, as computed by some similarity or distance measure. Two words are regarded as similar by this method if there is a large overlap in the two sets of documents in which they occur.

Notice that a document occurrence vector is substantially different from a word co-occurrence vector, in which each entry corresponds to a word in the lexicon rather than to a document. Notice also that a word co-occurrence vector that is constructed

from the relationship "joint occurrence in a document" represents exactly the same statistical information as a document occurrence vector. However, the information is represented differently in the two types of vectors, which yields different similarity values even if the same similarity metric is used. This chapter focuses on the word co-occurrence representation, though the similarity metrics of the following sections are relevant also for document occurrence vectors.

19.3. A schematic structure of similarity measures

Given representations of words as co-occurrence vectors, we need a mathematical measure for the degree of similarity between pairs of vectors. Either a similarity measure or a distance measure can be used: the value given by a similarity measure is proportional to the degree of similarity between the vectors (higher values for higher similarity) while the value of a distance measure is proportional to the degree of dissimilarity between the vectors (higher values for lower similarity).[1] In the rest of this chapter we refer to the two types of measures interchangeably, making the distinction only when necessary. This section describes, in a schematic way, the major statistical factors that are addressed by similarity and distance measures. The next section presents several concrete measures and relates their structure to the general scheme.

Similarity measures are functions that compute a similarity value for a pair of words, based on their co-occurrence vectors and some additional statistics. These functions typically consist of several components that quantify different aspects of the statistical data. The following subsections specify the major statistical aspects, or factors, which are addressed by similarity measures.

19.3.1. Word-attribute association

The first aspect to be quantified is the degree of **association** between a word *u* and each of its attributes, denoted by *assoc(u,att)*. The simplest way to quantify the degree of association between *u* and *att* is to set *assoc(u,att)= freq(u,att*), which is the original value in the vector entry corresponding to *att*. This measure is too crude, though, since it over-emphasizes the effect of words and attributes that are frequent in the corpus and are therefore a-priori more likely to co-occur together. Most similarity measures use more complex definitions for *assoc(u,att)* which normalize for word frequencies and possibly scale the association value according to individual attribute weights.[2]

It should be noted that mathematical definitions for word association were developed also for other applications such as extracting collocations and technical terms (see Chapter 7 of this book). In these tasks obtaining a good measurement for the degree of association is the goal of the system. In the context of word similarity, association measurement is only an intermediate step that appropriately scales the values of vector entries in order to facilitate their comparison with other vectors.

19.3.2. Joint association

The degree of similarity between two words *u* and *v* is determined by comparing their associations with each of the attributes in the two co-occurrence vectors. The two words will be regarded as similar if they tend to have similar degrees of association

[1] Our use of the term **distance measure** is not restricted to the mathematical notion of a **distance metric**, and is rather used in the general sense of denoting a measure of dissimilarity.

[2] When using a concrete similarity measure one might think of the values in the co-occurrence vector of *u* as denoting the value *assoc(u,att)*, for each attribute *att*. We prefer to consider *freq(u,att*) as the original value of a vector entry and *assoc(u,att)* as being part of the similarity formula. This view yields a general vector representation which summarizes the raw statistics of the word and may be fed into different similarity measures.

(strong or weak) with the same attributes. In order to capture such correspondence, similarity measures include a component that combines two corresponding association values, *assoc(u,att)* and *assoc(v,att)*, into a single value. We call this component the **joint association** of *u* and *v* with respect to the attribute *att,* and denote it by *joint(assoc(u,att),assoc(v,att))*. The *joint* operation is intended to compare the two values *assoc(u,att)* and *assoc(v,att)*, producing a high value when the two individual association values are similar (this is the case for similarity measures; for distance measures the joint association value is low when the two values are similar).

In some measures the comparison between the two association values is by the joint operation, like when defining the joint operation as the ratio or difference between the two values. In other measures the joint operation is defined by a different type of mathematical operations, like multiplication, that do not directly express a comparison. In these cases the comparison is expressed in a global manner, so that the normalized sum (see below) of all joint associations will be high if on average corresponding association values in the two vectors are similar. The joint association operation may also include a weighting factor for each attribute, which captures its "importance" with respect to similarity calculations.

19.3.3. Normalized sum of joint associations

The joint association of *u* and *v* with respect to the attribute *att* can be regarded as the contribution of the individual attribute to the similarity between *u* and *v*. The overall similarity between the two words is computed by summing the joint association values for all attributes. This sum should be normalized by a normalization factor, denoted by *norm*, whenever there is a variance in the "length" of co-occurrence vectors of different words. That is, different words may have a different number of active attributes (attributes with which they co-occur), as well as different association values with their attributes. Normalization of the sum is required to avoid scaling problems when comparing different word pairs. In some measures the normalization is implicit, since the *assoc* or *joint* operations already produce normalized values. In these cases the factor *norm* is omitted.

Using the three factors above, we now define a schematic form of similarity measures:

$$(2)\quad sim(u,v) = \frac{1}{norm}\sum_{att} joint\big(assoc(u,att), assoc(v,att)\big)$$

The schematic form illustrates the statistical "reasoning" captured by similarity measures. Distance measures have the same schematic form, where the *joint* operation produces higher values when the two association values are dissimilar. Many concrete similarity measures follow this schematic form closely. While some variation in the form of a similarity formula and its components may be possible, the statistical reasoning behind the formula still corresponds pretty much to the general schema.

19.4. Concrete similarity and distance measures

Word-attribute associations, joint associations and normalization may be captured by different mathematical formulas. This section presents several concrete measures that have appeared in the literature, analyzes their structure and illustrates their correspondence to the general schema (2). While quite a few alternative measures appear in the literature, rather little work has been done to compare them empirically and to analyze the effects of their components. Further research is therefore necessary to obtain a better understanding of the desired form of word-similarity measures. It should be noted that the measures described here give a range of different types of alternatives, though the list is not comprehensive.

19.4.1. "Min/Max" measures

By the name "Min/Max" measures we refer to a family of measures that have the following general form (3):

(3) $$sim(u,v) = \frac{\sum_{att} \min(assoc(u,att), assoc(v,att))}{\sum_{att} \max(assoc(u,att), assoc(v,att))}$$

This form may be regarded as a weighted version of the Tanimoto measure, also known as the Jaccard measure (see, for example, in Ruge (1992)). The Tanimoto measure assigns *assoc* values that are either 1, if the attribute is active, or 0, if it is inactive, while the Min/Max measure allows for any *assoc* value.

With respect to the schema (2), we can interpret the "min" operation as playing the role of the *joint* operation, while the sum of "max" values is the normalizing factor *norm*. This selection of *joint* and *norm* may be motivated by the following interpretation. Regard (4) as being composed of two components.

(4) *joint(assoc(u,att),assoc(v,att))*=min*(assoc(u,att),assoc(v,att))*

(5) $$\frac{\min(assoc(u,att), assoc(v,att)}{\max(assoc(u,att), assoc(v,att)}$$

(5) is the ratio which compares the two association values. This ratio ranges between 0 and 1, obtaining higher values when the two association values are closer to each other. The ratio is multiplied by a weighting factor, max*(assoc(u,att),assoc(v,att)),* which means that the effect of an attribute on the comparison of two words is determined by its maximal association value with one of the words. Thus, the *joint* operation is composed of the product of two components: the first measures the closeness of the two association values and the second weighs the importance of the attribute with respect to the given pair of words. Finally, to obtain the complete formula, the normalization factor *norm* is simply the sum of all attribute weights.

19.4.1.1. Log-frequency and global entropy weight

Given the general "Min/Max" form it is necessary to define the *assoc* operation. Grefenstette (1992, 1994) defines *assoc* as in (6), which was applied to co-occurrence vectors that are based on grammatical co-occurrence relations.

(6) $$assoc(u,att) = \log(freq(u,att) + 1) \cdot Gew(att)$$

where *frequ(u,att)* is the co-occurrence frequency of the word *u* with the attribute *att*. *Gew* stands for **Global-entropy-weight** for the attribute, which measures the general "importance" of the attribute in contributing to the similarity of pairs of words (7).

(7) $$Gew(att) = 1 - \frac{1}{\log nrels} \sum_{v} -P(v|att) \cdot \log(P(v|att))$$

where *v* ranges over all words in the corpus, *nrels* is the total number of co-occurrence relations that were extracted from the corpus, and the basis of the logarithm is 2. The sum $\sum_{v} -P(v|att) \cdot \log(P(v|att))$ is the entropy of the empirical probability distribution (as estimated from corpus data) *P(v|att)*, which is the probability of finding the word *v* in an arbitrary co-occurrence relation of the attribute *att*. The normalization value log *nrels* is an upper bound for this entropy, making the normalized entropy value $\frac{1}{\log nrels} \sum_{v} -P(v|att) \cdot \log(P(v|att))$ range between 0 and 1. This value, in proportion to the entropy itself, is high if *att* occurs with many different words and the distribution *P(v|att)* is quite uniform, which means that the occurrence of *att* with a particular word *v* is not very informative. *Gew* is inversely proportional to the normalized entropy value, obtaining high values for attributes which are relatively "selective" about the words with which they occur.

In summary, Grefenstette's association value is determined as the product of two factors: the absolute co-occurrence frequency of *u* and *att*, whose effect is moderated by the log function, and a global weight for the importance of the attribute. It should be noticed that the combination of these two factors balances (to some extent) the effect of attribute frequency. A frequent attribute in the corpus (corresponding to a frequent word *w*) is likely to occur more frequently with any other word, leading to

relatively high values for *freq(u,attt)*. This may amplify too much the effect of co-occurrence with frequent attributes. On the other hand, since a frequent attribute is likely to occur with many different words, it is also likely to have a relatively high entropy value for *P(v|att)*, yielding a relatively low *Gew* value. This effect of *Gew* may be compared with the commonly used IDF (Inverse Document Frequency) measure for attribute weight in IR, which is inversely proportional to the frequency of the attribute. The *Gew* measure was borrowed from the IR literature (Dumais 1991).

19.4.1.2. Mutual information

Dagan et al. (1993, 1995) have also used the "Min/Max" scheme for their similarity measure, which was applied to co-occurrence relations within a small window. They adopted a definition for *assoc* (8) which is based on the definition of mutual information in information theory (Cover & Thomas, 1991) and was proposed earlier for discovering associations between words (Church & Hanks, 1990; Hindle, 1990; see also Chapter 6 of this book).

$$(8)\ assoc(u,att) = \log \frac{P(u,att)}{P(u)P(att)} = \log \frac{P(att|u)}{P(att)} = \log \frac{P(u|att)}{P(u)}$$

Mutual information normalizes for attribute frequency, as can be seen in the conditional form of the formula. Its empirical estimation is sensitive to low-frequency events: due to data sparseness, the estimated value for *P(u,att)* may be too high for many word-attribute pairs that co-occur once or twice in the corpus, yielding an inflated association for these pairs. This is a well known problem when using mutual information to discover collocations or technical terminology, where identifying specific associations is the primary goal of the system. The problem is less severe when the measure is used to quantify many associations simultaneously as the basis for comparing two co-occurrence vectors.

19.4.2. Probabilistic measures

The measures described in this subsection are comparative probability distributions. In the context of computing word similarity, each word is represented by a probability distribution constructed by defining *assoc(u,att)=P(att|u)*. That is, the word *u* is represented by the conditional probability distribution which specifies the probability for the occurrence of each attribute given the occurrence of *u*. The similarity (or distance) between two words *u* and *v* is then measured as the similarity (or distance) between the two conditional distributions *P(att|u)* and *P(att|v)*.

19.4.2.1. KL divergence

The **Kullback-Leibler (KL) divergence**, also called **relative entropy** *(*Cover & Thomas, 1991), is a standard information theoretic measure of the dissimilarity between two probability distributions. It was used for distributional word clustering (Pereira et al., 1993) and for similarity-based estimation in (Dagan et al., 1994). When applied to word-attribute co-occurrence it takes the following form (9).

$$(9)\ D(u||v) = \sum_{att} P(att|u) \cdot \log \frac{P(att|u)}{P(att|v)} \cdot \log \frac{P(att|u)}{P(att|v)}$$

The KL divergence is a non-symmetric measure, as in general $D(u||v) \neq D(v||u)$. Its value ranges between 0 and infinity, and is 0 only if the two distributions are identical. With respect to the schema (2), the *joint* operation is composed of two components: a log-scaled ratio which compares the two association values *P(att|u)* and *P(att|v)*, and a weighting factor *P(att|u)* for the contribution of *att* to *D(u||v)*.

D(u||v) is defined only if *P(att|v)* is greater than 0 whenever *P(att|u)* is. This condition does not hold in general when using the Maximum Likelihood Estimator (MLE), where the estimate for *P(att|v)* is 0 when *freq(att,v)=0*. This forces using a

smoothed estimator which assigns non-zero probabilities for all *P(att|v)* even when *freq(att,v)=0*. However, having zero association values for many word-attributes pairs gives a big computational advantage, which cannot be exploited when using the KL divergence as a dissimilarity measure. Furthermore, the need to use a smoothing method both complicates the implementation of the word similarity method and may introduce an unnecessary level of noise into the data. The problem is remedied by the symmetric measure of **total divergence to the average**, which is described next. KL-divergence was used for word similarity by Dagan et al. (1994), who later adopted the total divergence to the average.

19.4.2.2. Total divergence to the average

The total divergence to the average, also known as the **Jensen-Shannon divergence** (Rao, 1982; Lin, 1991), was used for measuring word similarity by Lee (1997) and by Dagan et al. (1997) and is defined as (10),

$$(10)\ A(u,v) = D(u||\frac{u+v}{2}) + D(v||\frac{u+v}{2})$$

where $\frac{u+v}{2}$ is a shorthand (11).

$$(11)\ \frac{u+v}{2} = \frac{1}{2}\left(P(att|u) + P(att|v)\right)$$

A(u,v) is thus the sum of the two KL-divergence values between each of the distributions *P(att|u)* and *P(att|v)* and their average $\frac{u+v}{2}$. It can be shown (Lee, 1997; Dagan et al., 1997) that *A(u,v)* ranges between 0 and $2\log 2$.

By definition, $\frac{u+v}{2}$ is greater than 0 whenever either *P(att|u)* or *P(att|v)* is. Therefore, the total divergence to the average, unlike KL-divergence, does not impose any constraints on the input data. MLE estimates, including all zero estimates, can be used directly.

19.4.2.3. L_1 norm (taxi-cab distance)

The L_1 **norm**, also known as the **Manhattan distance**, is a measure of dissimilarity between probability distributions (12).

$$(12)\ L_1(u,v) = \sum_{att}\left|(P(att|u) - P(att|v))\right|$$

This measure was used for measuring word distance by Lee (1997) and Dagan et al. (1997). It was shown that *A(u,v)* ranges between 0 and 2. With respect to the schema (2), the *joint* operation compares the two corresponding association values *P(att|u)* and *P(att|v)* by measuring the absolute value of their difference. The use of a difference (rather than a ratio) implicitly gives higher importance to attributes where at least one of the two association values is high.

Empirical work, summarized in Dagan et al. (1997) and detailed in Lee (1997), compared the performance of total divergence to the average and L_1 norm (as well as **confusion probability** (Essen & Steinbiss, 1992)) on a pseudo-word-sense disambiguation task. Their method relies on similarity-based estimation of probabilities of previously unseen word co-occurrences. The results showed similar performance of the disambiguation method when based on the two dissimilarity measures, with slight advantage for the total divergence to the average measure.

19.4.3. The cosine measure

The **cosine measure** has been used extensively for IR within the framework of the vector space model (Frakes, 1992, chapter 14). It was applied also for computing word similarity, for example by Ruge (1992), who used co-occurrence vectors that are based on grammatical relations. The cosine measure is defined as (13).

$$(13)\quad \cos(u,v) = \frac{\sum_{att} assoc(u,att) \cdot assoc(v,att)}{\sqrt{\sum_{att} assoc(w_1,att)^2} \cdot \sqrt{\sum_{att} assoc(w_2,att)^2}} .$$

This is the scalar of product of the normalized, unit-length, vectors produced by the *assoc* values. With respect to the schema (2), the *joint* operation can be interpreted as the product of the two corresponding association values while the *norm* factor is the product of the lengths of the two *assoc* vectors. Combined together, these two operations entail a high cosine value if corresponding *assoc* values are overall similar (obtaining the effect of a comparison), giving a higher impact for larger *assoc* values.

Given the general cosine form, it is necessary to define the *assoc* operation. Ruge (1992) found the definition *assoc(u,att)*=ln*(freq(u,att))* to be more effective than either setting *assoc(u,att)=freq(u,att)* or *assoc(u,att)=1 or 0* according to whether the attribute is active or inactive for *u*. In analogy to other measures, one might try adding a global weight for the attribute (like *Gew*) or some normalization for its frequency. Ruge tried a somewhat different attribute weight, which obtains a low value when attribute frequency is either very high or very low. However, this global attribute weighting did not improve the quality of the similarity measure in her experiments.

19.5. Examples

This section gives illustrative examples for word similarities identified by the methods described in this chapter. The statistical data was collected from a sample of about 10,000 Reuters articles, which were parsed by a shallow parser. The co-occurrence vectors were based on statistics of grammatical co-occurrence relations, as described in Section 19.2.1.

Table 19.1 lists the most similar words for a sample of words, sorted by decreasing similarity score, as computed by four different similarity measures. Table 19.2 provides data about common context words that contributed mostly to the similarity scores of given pairs of words. For example, the first line of Table 19.2 specifies that the words *last, next* etc. co-occurred, as modifying adjectives, with both nouns *month* and *week*.

Tables 19.1 and 19.2 HERE

As can be seen from the tables, many of the similarities identified indeed correspond to meaningful semantic similarities, yet some degree of noise exists (the examples chosen are relatively "clean", other examples may include a higher level of noise). These results indicate that there is still a lot of room for future research that will improve the quality of statistical similarity measures and reduce the amount of noise in their output.

19.6. Conclusions

Automatic computation of various forms of word similarity has been recognized as an important practical goal. Word similarities are useful for several applications, for which the statistical computation may be used either in a fully automatic manner or in a semi-automatic way, letting humans correct the output of the statistical measure. Substantial research work in this area has

resulted in several common principles and a variety of concrete methods. Yet, methods of this type are very rarely integrated in operational systems. Further research is required to establish "common practices" that are known to be sufficiently reliable and useful for different applications.

Acknowledgments

The writing of this chapter benefited from the knowledge and results acquired in joint work with colleagues on statistical word similarity. The measure of Section 19.4.1.2 was developed in joint work with Shaul Marcus and Shaul Markovitch. The probabilistic measures of Section 19.4.2 were proposed for the purpose of measuring word similarity in joint work with Lillian Lee and Fernando Pereira. The scheme of Section 19.3 and the examples of Section 19.5 were obtained in joint work with Erez Lotan at Bar Ilan University.

References

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. Word sense disambiguation using statistical methods. In *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, pages 264-270.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and Lexicography. *Computational Linguistics*, 16, pages 22-29.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. New York: John Wiley.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pages 56--63.

Dagan, Ido, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pages 164--171.

Dagan, Ido, Shaul Marcus, and Shaul Markovitch. 1995. Contextual word similarity and estimation from sparse data. *Computer Speech and Language*, 9:123--152.

Dagan, Ido, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, pages 272--278.

Dumais, Susan T.. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments & Computers* 23:229-236.

Essen, Ute and Volker Steinbiss. 1992. Co-occurrence smoothing for stochastic language modeling. In *Proceedings of ICASSP*, volume 1, pages 161--164.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930-1955. In Philological Society, editor, *Studies in Linguistic Analysis*. Blackwell, Oxford, pages 1-32. Reprinted in *Selected Papers of J. R. Firth*, edited by F. Palmer. Longman, 1968.

Gale, William, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26: 415-439.

Grefenstette, Gregory. 1992. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, Copenhagen, Denmark, 89-97.

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

Grishman, R., L. Hirschman, and Ngo Thanh Nhan. 1986. Discovery procedures for sublanguage selectional patterns - initial experiments. *Computational Linguistics*, 12:205-214.

Grishman, Ralph and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Proceedings of DARPA Conference on Human Language Technology*, San Francisco, California, 254-259.

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

Hindle, D. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pa., pages 268-275.

Jelinek, Frederick, Robert L. Mercer, and Salim Roukos. 1992. Principles of Lexical Language Modeling for Speech Recognition. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*, New York: Marcell Dekker, Inc., 651-699.

Jiang, Jay J. and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING)*, Taiwan.

Karov, Yael and Shimon Edelman. 1996. Learning similarity-based word sense disambiguation from sparse data. In *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, 42-55.

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

Lin, Dekang. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pages 64--71.

Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transaction Information Theory*, 37:145-151.

Martin, W.J.R., B.P.F. Al, and P.J.G. van Sterkenburg. 1983. On the processing of text corpus: from textual data to lexicographical information. In R.R.K. Hartman (ed.) *Lexicography: Principles and Practice.* Academic Press, London.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pages 183-190.

Rao, C. Radhakrishna. 1982. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankyha: The Indian Journal of Statistics*, 44(A):1-22.

Resnik, Philip. 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, Menlo Park, California, 56-64.

Resnik, Philip. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Mass., 54--68

Ruge, Gerda. 1992. Experiments on linguistically-based term associations. *Information Processing & Management*, 28:317-332.

Schutze, Hinrich. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, Minneapolis, MN, 787-796.

Schutze, Hinrich. 1993. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, eds, *Advances in Neural Information Processing Systems 5*, Morgan Kaufman, San Mateo, California, 895-902.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19: 143-177.

Word	Measure	Most similar words
month	Average-KL	year week day quarter night December thanks
	Cosine	year week day quarter period dlrs January February
	Minmax-Gew	week year day quarter period February March sale
	Minmax-MI	week year day January February quarter March
shipment	Average-KL	delivery sale contract output price export registration
	Cosine	delivery contrasale output price
	Minmax-Gew	delivery output contract surplus period registration
	MI	delivery output contract registration surplus election
copper	Average-KL	Rubber nickel LPG Opec composite sorghum Issue
	Cosine	rubber composite LPG truck issue Opec sorghum
	Minmax-Gew	nickel zinc cathode gold coffee Jeep silver spot
	Minmax-MI	nickel zinc cathode coffee Jeep coal semiconductor
boost	Average-KL	accommodate curb increase reduce depress keep
	Cosine	accommodate depress curb increase cut translate reduce
	Minmax-Gew	reduce keep curb cut push increase limit exceed hit
	Minmax-MI	curb keep reduce hit push increase exceed cut limit
accept	Average-KL	reject submit formulate disseminate approve consider
	Cosine	disseminate formulate reject submit unrestricted
	Minmax-Gew	reject approve submit consider threaten adjust extend
	Minmax-MI	submit threaten approve reject reaffirm consider pass

Table 19.1 Sample of most similar words by four different similarity measures: total divergence to the average, cosine measure with *assoc(u,att)*=ln*(freq(u,att)),* Min/Max with log-frequency and global entropy weight.

Word-1	Word-2	Rel	Common context words
Month	week	a-n	last next few previous past several
		n-n	bill period
		n-prep-n	end sale export
	year	a-n	last next first previous past few recent several
		n-n	period
		n-prep-n	end dlrs export
shipment	delivery	n-n	April June March January
		n-prep-n	cent dlrs June April January year
		n-prep-n	tonne corn
	Export	n-n	coffee corn June January grain
		n-prep-n	tonne year January Europe
		s-v	rise fall
		n-prep-n	tonne tariff sugar
agency	news	n-n	official Tass Tanjug Association APS SPK MTI
	ministry	s-v	say announce add
		n-n	statement source official
	department	s-v	say report note assign point announce
		n-n	statement plan official
law	regulation	n-n	Banking
		a-n	federal new German current
		n-prep-n	accordance change
		s-v	allow
		v-o	change
	legislation	n-n	trade reform preference
		a-n	special new recent
		n-prep-n	change year
		s-v	require call

Table 19.2 Common context words, listed by syntactic relation, which were major contributors to the similarity score of Word-1 and Word-2. Key: **n-n** noun-noun (noun compound), **a-n** adjective-noun, **n-prep-n** noun-preposition-noun, **s-v** subject-verb, **v-o** verb-object. The underscored position in the relation corresponds to the position of each of the two similar words (Word-1/2) while the other position corresponds to the common context word.