9-5-2013

Mismatches between Humans and Latent Semantic Analysis in Document Similarity Judgments

Kyunghun Jung

Kyunghun Jung
Candidate

Psychology
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Eric Ruthruff, Chairperson

Timothy Goldsmith, Co-Chair

Harold Delaney

George Luger

MISMATCHES BETWEEN HUMANS AND

LATENT SEMANTIC ANALYSIS

IN DOCUMENT SIMILARITY JUDGMENTS

by

KYUNGHUN JUNG

B.A., Psychology, Korea University, 2004
M.A., Experimental Psychology, Korea University, 2006

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Psychology

The University of New Mexico
Albuquerque, New Mexico

July, 2013

DEDICATION

To my wife Soonae: Thank you for loving me.

ACKNOWLEDGMENTS

I also want to say thank you to a few more people who helped me in school and in life.

Dr. Eric Ruthruff, my advisor and dissertation chair, invested so much energy into this project. Because of him, I was able to start my new life here at UNM, which led to other amazing events.

Dr. Tim Goldsmith, my advisor and co-chair, always smiled at me and made me feel comfortable studying as a foreign student. I appreciate your support for this project and all the kindness you showed me.

Dr. Harold Delaney demonstrated a scholarly standard that I will pursue during my whole life.

Dr. George Luger generously agreed to be a committee member for this dissertation, which made my graduation possible.

Dr. Lee and Dr. Pincombe generously provided their data that played a critical role in this dissertation project.

Finally, Soonae, I love you from the bottom of my heart. You are teaching me something I have never learned from anyone else.

Mismatches between Humans and Latent Semantic Analysis in Document Similarity Judgments

by

Kyunghun Jung

B.A., Psychology, Korea University, 2004
M.A., Experimental Psychology, Korea University, 2006
Ph.D., Psychology, University of New Mexico, 2013

ABSTRACT

Modeling how humans judge the semantic similarity between documents (e.g., abstracts from two different psychology articles) is an interesting and challenging topic in cognitive psychology. It also has practical implications for developing artificial intelligence (AI) systems, especially those designed for retrieving relevant information from a large database in response to a given query (e.g., finding new research articles related to a given abstract). Conversely, AI algorithms can provide a useful tool for testing human cognitive models. They can precisely simulate the consequences of specific assumptions about cognition, and these consequences can then be compared against actual human performance. In the process of developing both human cognitive models and AI models, investigating the discrepancy between human and AI performance is essential, although it has rarely been explored with respect to document relatedness judgments. In the current study, I identified a set of document pairs whose relatedness was judged radically differently between humans and a computational model called latent semantic analysis (LSA). Based on an examination of those misjudged document pairs, I proposed a tentative model of

human document relatedness judgment, called the key-features overlap model. According to this model, document relatedness judgments by humans and computational algorithms can be explained, in part, by the degree of word-pair association across documents. Critically, it suggests that, to judge document relatedness, humans focus primarily on the association between the keywords in each document, while computational algorithms including LSA typically do not. Modifications of target documents to emphasize their keywords, while also providing keyword-relevant background documents to LSA improved LSA's document relatedness judgments. Such improvement demonstrated the usefulness of the key-feature overlap model-based approach for improving AI algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

The overarching goal of the current study is to contribute to a better understanding of human cognition, and apply the psychological knowledge of human cognition to a selected class of artificial intelligence (AI) algorithms, so that the algorithms can better mimic human behavior. Among various cognitive processes, the current study is focused on how humans judge the semantic similarity between text documents, especially paragraphs. This specific cognitive task is essential to many professions in this information-overload era (Toffler, 1970), where information is often summarized in a single paragraph. As an example, to find articles relevant to the target domain out of a large database, researchers often glance over abstracts of articles and judge their relevance. This process is feasible for a human if the number of possible documents is small, but it is highly impractical when the number of documents to be searched is enormous, as is typically the case[1].

To assist humans with time- and effort-consuming document relatedness judgments, computer scientists have developed AI algorithms. As an example, Google Scholar recently started a service that searches for articles that match a researcher's profile (Connor, 2012). The critical component of the service is AI algorithms that determine semantic relatedness between pieces of textual information. A goal of the current study is to refine human models of document relatedness judgments, and utilize the human model to improve a class of AI algorithms.

Luger (2009) defined AI as "the study of the mechanisms underlying intelligent behavior through the construction and evaluation of artifacts designed to enact those mechanisms" (p. 675). With regard to the construction of artifacts such as algorithms, there are two main approaches (Ashby, 1960; McCorduck, 2004). In the first approach, AI researchers build the

[1] As of February 2010, there were 1.7 million full-text, peer-reviewed biomedical and life sciences articles in an UK electronic journal article archive, PubMed Central (Kiley, 2010).

artifacts without any intention of modeling how humans perform a task. Just as aeronautical technologies based on optimizing performance in wind tunnels (rather than mimicking birds) yielded success with aircraft flights, AI algorithms that are not intended to model human intelligence can nevertheless yield successful performance (Russell & Norvig, 2003). However, there is an alternative approach that has inspired many AI researchers, as well as the current study. In this approach, AI researchers design their algorithms to mimic human intelligent behavior and, if they fail, to do so in the same way that human intelligence fails. This latter approach may be effective, especially when humans outperform current AI algorithms on a task of interest. For example, with respect to facial recognition, the human visual system shows good performance across a wide range of environmental conditions, outperforming even the most cutting-edge AI algorithms. In this case, development of facial recognition algorithms can be inspired by (psychological) research about human facial recognition processes (Sinha, Balas, Ostrovsky, & Russell, 2006).

In the current study, I consider the averaged semantic similarity ratings of document pairs made by humans as a "correct" indicator of the degree of relatedness between documents. Therefore, AI algorithms producing different document relatedness ratings from humans are assumed to be inaccurate. With these assumptions, I attempt to understand the mechanisms underlying human judgments of document relatedness, and explain why certain AI algorithms make different judgments from humans. To do this, I will first try to explain human document relatedness judgments based on one traditional psychological model of human similarity, Tversky's (1977) contrast model. Then, I will compare document relatedness judgments of humans and an AI algorithm called latent semantic analysis (LSA).

The current study has the following structure. In Chapter 2, I introduce psychological models of human similarity judgments, especially Tversky's (1977) contrast model, and an algorithmic approach of computing document similarity, LSA. In this chapter, I also introduce some critical studies that examined the degree of agreement between humans and LSA with respect to their document relatedness judgments. To compare the document relatedness judgments of humans and LSA under LSA's optimal parameter setting, in Chapter 3, I test various parameters of LSA, and define an optimal set of parameters, yielding the highest degree of agreement between humans and LSA. Using these optimal parameters, in Chapter 4, I identify some problematic document pairs whose relatedness is judged radically differently between humans and LSA, called misjudged document pairs. In Chapter 5, to explain human document relatedness judgments of those misjudged document pairs, I first evaluate the contrast model of human document relatedness judgments. As will be shown later, the contrast model does not provide a good explanation of human performance, suggesting that it requires modification. In Chapter 6, I propose such a modification called the key-features overlap model. I also report a human experiment to test whether this new model can better explain human document relatedness judgments than its predecessor. In Chapter 7, I modify documents subject to LSA so that LSA can judge their relatedness in a more similar way to the key-features overlap model, which shows promising results. Finally, Chapter 8 summarizes the previous chapters and discusses the implications of the current study.

Chapter 2. Background

In Chapter 2, I introduce background information related to the current study. First, I introduce psychological models of human similarity judgments. Those models are not specific to the semantic similarity judgments of documents, but, rather, attempt to explain the basis of human similarity judgments in general. Then, I introduce LSA, especially its underlying mechanisms of computing semantic similarity between documents. Finally, I introduce some previous studies that examined the level of agreement between humans and LSA with respect to their document relatedness judgments.

Psychological Models of Human Similarity Judgments

Similarity is a central theoretical construct in cognitive psychology. It has been employed to explain various cognitive processes such as generalization of learning to other stimuli (Osgood, 1949), categorization (Nosofsky, 1986), context effects on information retrieval (Roediger, 1990), and so on. The fact that similarity has been applied to various cognitive phenomena implies that it is a wide-ranging concept. Accordingly, there have been several different theoretical approaches to similarity, such as Tversky's contrast model (1977), Nosofsky's selective attention model (1986), and Shepard's geometric model (1962, 1987). Among the different approaches, Tversky's model is one of the most fundamental models, and shows large conceptual overlap with other models (Edelman & Shahbazi, 2012). This specific model forms the theoretical background of the current study, and thus is further introduced below.

The critical assumption of the contrast model is that objects are represented as a collection of individual features, and similarity judgments are basically a comparison of the commonalities in these features. For example, in his seminal work, Tversky (1977) focused on

isolated features of objects to account for object representation and perceived similarity. He assumed that objects are similar to each other to the extent that they share the same features, and dissimilar to the extent that features occur in one object but not the other. The similarity between two objects (A and B) can also be described as follows (also known as a ratio measure):

$$\frac{number\ of\ common\ features\ in\ A\ and\ B}{number\ of\ common\ features\ in\ A\ and\ B\ +\ number\ of\ unique\ features\ in\ A + number\ of\ unique\ features\ in\ B}$$

According to the contrast model, semantic similarity between documents can be also explained by the degree of overlap of individual features, such as syllables, words, or sentences. Regarding the unit of features, Tversky (1977) mentioned that "the features may correspond to components such as eyes or mouth [in comparing pictures of faces]; they may represent concrete properties such as size or color, and they may reflect abstract attributes such as quality or complexity" (p. 329), implying that there can be various units for feature representation. In the current study, I used the word as the unit of features following the bag-of-words model. The bag-of-words model regards a document as a collection of words, ignoring the word order. This is a widely-accepted model in computer science, and to provide compatible results to the specific research domain, I used the unit of word for the feature representation of documents.

The contrast model has been supported by many studies using simple stimuli such as line drawings or animal names. However, studies using more semantically complex stimuli, such as metaphors or analogies, have shown that the contrast model does not provide a good explanation of human semantic similarity judgments. For example, Gentner (1983) argued that interpreting an analogy (e.g., "an electric battery is like a reservoir") is fundamentally a search for a common relational structure between phrases or sentences, rather than a simple comparison of features.

Features in an analogy are placed in correspondence on the basis of holding like roles in the relational structure, not on the basis of their intrinsic characteristics such as actual size or color. According to Gentner, the key to analogy comprehension is common systems of relations rather than the sheer number of matching features.

I should note, however, that Gentner's (1983) criticism of the contrast model may be limited to the case of analogy or metaphor comprehension. In other words, the contrast model could still be viable with regard to explaining human semantic similarity judgments of *plain texts*. Even if the contrast model provides a naive explanation only (i.e., document similarity is based on word-pair similarity), such a simple explanation can provide a basic framework with which one can approach the complex human cognitive processes of document relatedness judgments. However, as one can expect, the contrast model may need some modification to satisfactorily explain human document relatedness judgments. One major goal of the current study, therefore, is to evaluate the contrast model for document relatedness judgments and identify possible modifications of the model. To obtain insights for such modification, I compare document relatedness judgments between humans and a computer algorithm, LSA, which makes its document relatedness judgments primarily based on the degree of overlap of words (i.e., a sort of simulation of the contrast model). Differences between humans and LSA with respect to their document relatedness judgments will indicate the weaknesses of not only LSA but also the contrast model, and may suggest how one should modify the current contrast model to better explain human document relatedness judgments.

LSA

In computer science, numerically representing documents and measuring their semantic relatedness have been widely studied. The vector-space model (Salton & McGill, 1983) is one

of the most popular document representation models, which also provides the basis of LSA. In this model, documents are first transformed into a *word-by-document matrix*. Rows of the matrix correspond to the words observed across the documents, and the columns correspond to individual documents. The cell values are the frequencies of words within each document. These cell values can be weighted in two respects (Polettini, 2004): to what degree a word is important in representing a document's topic (*local weighting*), and to what degree a word is important in distinguishing one document from another according to their topics (*global weighting*). Using the (weighted) cell values, each document can be represented as a vector, V_j, on a multidimensional *word-space*, where each dimension corresponds to each word, as below:

$$V_j = [w_{1j}, w_{2j}, \ldots w_{ij}, \ldots w_{tj}]^T$$

where V_j represents the j^{th} document's vector on a t dimensional word-space, w_{ij} represents the (weighted) frequency of the i^{th} word in the j^{th} document, and t is the total number of words.

One critical assumption of the vector-space model is that documents describing the same topic have similar words. According to this assumption, documents with similar topics tend to have more overlap in their vectors, and be positioned close to each other in the word-space, while ones with dissimilar topics tend to be separated. In sum, the geographic relationship between document vectors can be used to measure the semantic similarity of documents (see Figure 1). Notably, there is a close correspondence between the contrast model's explanation of document similarity and the way a vector-space model decides document similarity (i.e., overlap of words).

Cosine similarity, which is based on the cosine of the angle between two document vectors, is widely used to measure the distance between two document vectors' orientation. A cosine similarity close to 1 indicates high semantic similarity between documents, and a value close to 0 indicates unrelated documents (Andrews & Fox, 2007; Salton & McGill, 1983).

$$cosine\ similarity\ (V_j, V_k) = \frac{\sum_{i=1}^{t}(V_{ij} \times V_{ik})}{\sqrt{\sum_{i=1}^{t} V_{ij}^2} \times \sqrt{\sum_{i=1}^{t} V_{ik}^2}}$$

where V_{ij} represents the frequency of the i^{th} word in the j^{th} document.

A word-by-document matrix usually has extremely high dimensionality corresponding to the number of words in a corpus, even after pre-processing[2]. Such high dimensionality yields high computational cost. Therefore, *dimension reduction* is typically applied to the word-by-document matrix (Sebastiani, 2002) before documents are represented in the multidimensional space. After the dimension reduction, document vectors are represented on a subspace with a lower-dimensionality, *k*, which is chosen by the user.

One advantage of this dimension reduction is computational efficiency. However, it also causes loss of information relative to that contained in the original matrix. At first glance, this seems to be detrimental to accurately representing documents. However, it can actually increase the accuracy of the document representation, with the explanation being that noise is removed from the original matrix through the dimension reduction process. There are several different

[2] Articles (e.g., 'the', or 'an'), pronouns (e.g., 'he', or 'she'), prepositions (e.g., 'from', or 'to'), and conjunctions (e.g., 'and', or 'but'), which are called *stopwords,* occur too frequently across documents. Also, they have an approximately even probability of occurring across different topics of documents. As a result, they provide no information with respect to distinguishing documents according to their topics. They are excluded from the word-by-document matrix in the initial stage of pre-processing that also consists of several other steps. Appendix A summarizes the characteristics of various pre-processing steps.

dimension reduction techniques with such beneficial effects, such as singular value decomposition (SVD), non-negative matrix factorization (NNMF, Lee & Seung, 1999; 2001), random indexing (Sahlgren, 2005), and so on. NNMF is a matrix factorization technique similar to SVD which is shown in detail below, with a special restriction of not allowing negative values in the sub-matrices factorized from the original matrix. Random indexing is a dimension reduction technique which does not use matrix factorization. The basic idea of random indexing is that we can reduce the dimensionality of an original matrix by projecting it on a lower multidimensional subspace. By squeezing the original matrix into a lower dimensional space, we would lose some information and cause distortions of the original data. However, Sahlgren showed that even with the information loss and distortion, the newly represented data in the subspace approximately retains the spatial relationship across data points (documents) contained in the original matrix. Arguably, applying the above mentioned dimension reduction techniques to an original word-by-document matrix yields better document representation, because the new document representation retains the core relationship between documents with reduced noise obstructing the true relationship of documents.

Among the different dimension reduction techniques, applying SVD to the original word-by-document matrix is called latent semantic analysis (LSA). In the current study, I chose LSA as a representative AI algorithm of performing linguistic tasks, since LSA is well known for its ability to emulate human linguistic performance as shown at the end of this section. Below, I introduce SVD further.

SVD is a matrix factorization method that decomposes the original matrix into three sub-matrices as follows (Madsen, Hansen, & Winther, 2003),

$$A = USV^T$$

where A is a w X d matrix (the word-by-document matrix) with a rank of r, U is a unitary[3] w X r matrix (word-by-dimension matrix), S is an r X r diagonal matrix with non-negative real numbers on its diagonal (singular value matrix), and V^T is a unitary r X d matrix (dimension-by-document matrix).

Factorization of the original matrix (A) into this format is called standard or full SVD of A. This factorization exists for any matrix A. By multiplying these three sub matrices (U, S, and V^T), one can recover the original matrix (see the upper panel of Figure 2).

To achieve the dimension reduction, after the full SVD, one can intentionally discard small singular values located in the lower right corner of S by changing them to zero (dashed line inside of the S in the upper panel of Figure 2), while preserving the first k largest singular values (the solid line in the S). The corresponding columns of U, and the corresponding rows of V^T are discarded as well (see the lower panel of Figure 2). This is called reduced or truncated SVD, which is described as follows,

$$\hat{A} = U'S'(V^T)'$$

where U' is a w X k matrix whose columns are the first k columns of U, S' is a k X k diagonal matrix whose diagonal elements are the k largest singular values of S, and $(V^T)'$ is a k X d matrix whose rows are the first k rows of V^T.

[3] A square matrix M that meets the following is unitary, M'XM = MXM' = I, where M' is a transpose of M and I is the identity matrix.

By multiplying these three reduced sub-matrices (U', S' and (V^T)'), one can obtain $\hat{A}$ which is the best approximation of A (Berry, Dumais, & O'Brien, 1995).

After the reduced SVD on A, one can represent the documents as vectors on a k dimensional *singular-value-space*, which has k orthogonal axes. The axes are constructed so that the first axis explains the largest amount of variance of A, and the second axis explains the second largest amount of variance of A, and so on. Document vectors on this singular-value-space can be obtained by multiplying (V)' by S'. For example, if we set k = 2 (i.e., selecting the first 2 singular values in S), then (V)' is a d (number of documents) X 2 matrix, and S' is a 2 X 2 matrix. Accordingly, (V)' X S' will make another d X 2 matrix where the two cell values of each row are the coordinates of each document vector represented on the 2 dimensional singular-value-space.

Furnas et al. (1988) was the first to apply SVD to the vector-space model of documents. This method was later called *latent semantic analysis* by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), who showed better information retrieval performance by LSA than by traditional word-matching methods (see also, Dumais, 1991). Deerwester et al. argued that SVD uncovers latent semantic relations across documents that were buried in the original corpus, by capturing the most important information of the original word-by-document matrix.

Somewhat more surprising about LSA than its outstanding information retrieval performance is how well it models human linguistic behavior on various tasks. For example, it can simulate non-native English speakers' TOEFL (Test of English as a Foreign Language) test-taking performance, as well as schoolchildren's vocabulary learning (Landauer & Dumais, 1997). More directly related to the study, there is substantial agreement between humans and LSA with respect to their document relatedness judgments, as will be discussed below. However, there is

also a robust mismatch between humans and LSA with respect to their document relatedness judgments.

Agreement in Document Relatedness Judgments between Humans and LSA

A recent study examined the degree of agreement between humans and LSA with respect to their document relatedness judgments (Pincombe, 2004). Pincombe collected 50 news articles from the Australian Broadcasting Corporation's news mail service, and formed 1,225 pairs. Each news article had a single paragraph with a length of 51 to 126 words (median of 78.5). Eighty-three university students rated relatedness of those news article pairs using a five-point scale (one indicating "highly unrelated" and five indicating "highly related"). He also obtained LSA-cosine scores for the same pairs. The correlation between human ratings and cosine scores across the 1,225 document pairs was 0.60. Notably, this correlation was almost the same as the inter-rater reliability of 0.61 obtained from the 83 human readers. That is, LSA showed similar performance to a randomly selected human participant, in terms of the degree of agreement. However, as will be discussed in detail in Chapter 4, LSA makes radically different judgments from the averaged human ratings for certain document pairs. Such document pairs suggest that, despite the high level of agreement between humans and LSA, there is a systematic difference between them with respect to the document relatedness judgment, which will be intensively studied in the current research.

Regarding the systematic difference of document relatedness judgments, Lee, Pincombe, and Welsh (2005) proposed an interesting hypothesis. In their study, using the same 1,225 news article pairs, they tested the document relatedness judgment ability of a vector-space model with a specific similarity measure (common-features model, which is similar to the Tversky's ratio measure). They observed a correspondence pattern between humans and the model, as shown in

Figure 3. As shown in the upper left corner of Figure 3, there were numerous document pairs that were judged to be related by humans but not by the model. Based on this observation, Lee et al. hypothesized that document pairs judged to be related by humans but not by computational models would be the main source of the mismatch between humans and models, including LSA. However, this hypothesis has not been empirically tested yet. This is a major oversight because the pattern of mismatches between humans and computational models can reveal the strengths of humans and weaknesses of the models, which can be useful for building more complete human models and more accurate computational models of document relatedness judgments. Considering the correspondence between the contrast model and the vector-space model as described above, by comparing human and a variant of the vector-space model such as LSA with respect to their document relatedness judgments, it may be possible to understand to what degree human document relatedness judgments can be explained by the contrast model (simulated by LSA), and to obtain insights regarding how one should modify the vector-space model to make it better emulate human behavior.

Lee et al. (2005) and Pincombe (2004) generously provided the 50 news articles, as well as the human ratings for the 1,225 article pairs obtained in their studies. In the following two chapters, using that data, I identified the document pairs that were judged radically differently by humans and LSA. For ease of discussion, I will henceforth refer to such document pairs as the *misjudged document pairs*.

Chapter 3. Optimal Parameters of LSA

To examine differences between humans and computational models with respect to their document relatedness judgments, I will begin by studying document pairs whose relatedness is judged radically differently by humans and LSA. My rationale is that such extreme failures (i.e., the misjudged document pairs) may provide the clearest hints regarding flaws in computational models. With this rationale, the overarching goal of chapters 3 and 4 is to identify the misjudged document pairs from the 1,225 news article pairs. One could determine the misjudged document pairs by comparing humans' relatedness ratings and LSA-cosine scores obtained in the previous study (Pincombe, 2004). Although this seems a reasonable and simple approach, as described in Chapter 2, there are various LSA parameters such as local and global weighting schemes that can radically change LSA's document representations. Therefore, depending on the parameter setting chosen for the LSA procedure, one may obtain different LSA-cosine scores, and as a result, different sets of misjudged document pairs.

Among possible sets of the misjudged document pairs, a set determined based on an ideal parameter combination that yields the highest level of agreement between humans and LSA would be more informative than ones determined based on relatively poor parameter combinations. The purpose of Chapter 3, therefore, was to test various LSA parameters and find the combination of parameters that yielded the highest level of agreement between humans and LSA across the 1,225 of document pairs, which I call the *optimal parameters*. Specifically, I tested the following parameters, which are known to affect LSA's document representation significantly: (a) local and global weighting schemes (Lintean, Moldovan, Rus, & Mcnamara, 2010; Nakov, Popova, & Mateev, 2001), (b) dimensionality of the reduced SVD (Dumais, 1991;

Landauer & Dumais, 1997), and (c) the number of background documents (Bullinaria & Levy, 2007).

Based on a pilot study testing various weighting schemes (not reported here), I selected nine promising weighting schemes, including the log-entropy (Dumais, 1991), which showed superior performance to other weighting schemes. Regarding dimensionality, Dumais (2003) described the importance of choosing appropriate dimensionality as follows: "With too few dimensions, LSA performance is poor because there is not enough representational richness. With too many dimensions, performance decreases because LSA models the noise in the data thus reducing generalization accuracy" (p. 497). Based on some researchers' arguments for the importance of maintaining 300 dimensions (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), in the current study, I tested 19 different numbers of dimensions, distributed mostly between 200 and 300 (see the Method section for more information). Finally, I manipulated background documents. Background documents are defined as documents included in the original corpus subject to LSA along with the target documents, but whose representation or cosine similarity are not the main focus of the users who run LSA. Background documents are employed for constructing the multidimensional space in which target documents (in this case, the 50 news articles) are represented. Previous studies have shown that both size and quality (i.e., relevance to target documents) of the background documents affect LSA's document representations (Bullinaria & Levy, 2006). Especially, it is generally believed that LSA's performance improves as the number of background documents increases (Landauer & Dumais, 1997). In his critical study, Pincombe (2004) used only 314 news articles as background documents. In the current study, I employed 4,172 news articles as background documents as well as the 314 news articles used in Pincombe. Those 4,172 additional

documents came from the same source the 50 target news articles (the Australian news corpus), and were provided by the previous researchers (Lee et al., 2005; Pincombe, 2004).

In sum, I manipulated weighting schemes, dimensionality, and the number background documents to obtain the optimal parameters, defined as yielding the highest level of agreement between humans and LSA.

Method

Stimuli and Apparatus. The same 50 news articles used in Pincombe (2004) were used as the main stimuli (see Appendix B). They cover a variety of news topics, such as terrorism and hunger in Africa. All the background documents used in the current study came from the same source, the Australian news corpus, as the 50 target documents. Each background document has a single paragraph, with an average length of 154 words. The pre-processing and weighting schemes were implemented using a *Matlab* library called, the term to matrix generator (TMG; a MATLAB toolbox created by Zeimekis and Gallopoulus, 2012).

Procedure and Design. Stemming, normalization, and removal of stopwords and alphanumeric words are known to enhance LSA's performance, and so were implemented (Pincombe, 2004; Stone, Dennis, & Kwantes, 2011). Based on a pilot study, three local weightings (term frequency, logarithmic, and alternative logarithmic) and three global weightings (idf, entropy, and p-inverse) were selected as the most promising weighting schemes for improving the level of agreement between humans and LSA. Every possible combination of these local and global weighting schemes was tested.

To test the effects of the number of background documents, I compared the following three conditions: 0-, 314-, and 4,486-document background conditions. In the 0-document background condition, only the 50 target news articles were subjected to LSA without any

background documents. In the 314-document background condition, the same 314 news articles that were used in the previous study (Pincombe, 2004) were used as background documents (i.e., replication of the previous study). For the 4,486-document background condition, 4,172 news articles, along with the 314 documents were used as background.

Regarding the effect of dimensionality, the maximum possible dimensionality in each background condition is determined by the total number of documents used in LSA. For example, in the 0-document background condition the maximum dimension is 50 (i.e., the number of target documents). In this condition, I tested the five evenly distributed dimensions between 0 and 50: 10, 20, 30, 40 and 50. For the 314-document background condition, Pincombe (2004) observed relatively higher degrees of agreement between humans and LSA from the window of 200 and 300 dimensions. Therefore, I tested more narrowly distributed dimensions within this window as follows: 50, 100, 150, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 350, 360, and 364 (50 + 314 = 364 is the maximum dimension in this condition). In the 4,486-document condition, although the maximum dimension is 4,536 (50 + 4,486), I tested the same dimensions as in the 314-document condition because of the high computational cost of SVD at high dimensionality. Every possible combination of the nine weighting schemes and number of dimensions in each background condition was tested.

Results

Figure 4 shows the effect of the number of background documents, nine weighting schemes, and dimensionality on correlation between humans and LSA with respect to their document relatedness judgments. The greatest effect of parameter manipulation was observed from the manipulation of the number of background documents. The correlation coefficients from the 4,486-document condition (average $r = 0.67$) were higher than those from the 0-

document condition (average $r = 0.46$) or the 314-document condition (the average $r = 0.54$). Indeed, almost every parameter combination involving the largest number of background documents worked as well or better than every parameter combination involving the other sets of background documents. This result shows the beneficial effect of having numerous background documents on LSA's document relatedness judgments.

As dimensionality increased, the degree of agreement generally increased for the 0- and 314-document background conditions. However, in the 4,486-document background condition, the correlation started to slightly decrease after a dimensionality of around 300, forming an asymptote. The current result is consistent with the typical results of LSA, suggesting that there is not a single optimal dimensionality. That is, the optimal dimensionality depends on various other factors, such as the size of the original corpus.

Regarding the effect of weighting schemes, the local weighting showed a stronger effect on the agreement between humans and LSA than the global weighting. However, I should note that the overall effects of weighting schemes in the current study were relatively small compared to the effects of other parameters, especially background documents.

Notably, the highest correlation of 0.70 was obtained from the 4,486-document background condition at the dimensionality of 200, with the term frequency local and p-inverse global weighting scheme. I will henceforth refer to these as the optimal parameters. This correlation was significantly higher than the correlation of 0.60 observed by Pincombe (2004; $z = 4.25$, $p < .001$). Also, this correlation was significantly higher than the maximum correlation of 0.61 obtained from the simple vector-space model (not shown in Figure 4) without any dimension reduction. In other words, by applying SVD to the vector-space model (i.e., LSA) we can explain 12% more variance (49% – 37%) in human document relatedness ratings.

Discussion

The current results show that parameter settings can have profound effects on the level of agreement between humans and LSA. Using a large number of background documents was especially beneficial for LSA's performance, echoing the results from previous studies (e.g., Bullinaria & Levy, 2006). One may suspect that even more background documents would further increase the correlation (although I should note that the inter-rater agreement among human raters was only 0.61 for the target news article pairs, and therefore, it may be hard to further improve LSA's performance). However, to make a positive impact on LSA's performance, the background documents should not only be numerous but also relevant to the content of the target documents (Foltz, Britt, & Perfetti, 1994). There is a recent study tested the effect of various sets of background documents on LSA's document relatedness judgments (Stone, Dennis, & Kwantes, 2011). The researchers calculated the agreement between humans and LSA across the same 1,225 news article pairs as in the current study. Their background documents were the 55,021 Canada's *Toronto Star* newspaper articles (miscellaneous gossip paragraphs, from the year 2005), or 10,000 articles from the online encyclopedia, Wikipedia (http://www.wikipedia.org/)[4]. Surprisingly, the highest correlation between humans and LSA they obtained were about 0.10 and 0.40, respectively, from the gossip and Wikipedia background documents. These results suggest that having a large number of background documents does not necessarily produce good LSA performance, and great care is required in assembling the background documents.

There have been other attempts to increase LSA's document relatedness judgments by manipulating which background documents are utilized by LSA. So far, Gabrilovich and

[4] The researchers subjectively identified keywords of each of the 50 target news paragraphs. Then, using a high-performance text search engine, Lucence, they derived the 10,000 Wikipedia articles that are relevant to those keywords.

Markovitch (2007) showed the greatest success in terms of the correlation between humans and LSA (r = 0.72), using the same 1,225 news article pairs. Their background documents were 241,393 Wikipedia articles covering general topics. I should note that the current study's highest correlation of 0.70 is not significantly different from their highest correlation (z = 1.1, p = .27), even though the number of background documents in the current study was only 18.5% of that used by Gabrilovich and Markovitch. The current results suggest that, when target documents come from a specific corpus (e.g., Australian news articles), documents from the same corpus provide an especially good background, and thus even relatively small numbers of background documents can suffice to produce good LSA performance. However, I should also note that broad background documents may yield good LSA performance across various domains of target documents, whereas the same-source background documents may be well-suited only for the specific target documents.

Regarding the weighting schemes, the current study's optimal weighting scheme of term frequency local – p-inverse global did not support previous studies showing an advantage of using the log – entropy weighting scheme (e.g., Dumais, 1991; Nakov et al., 2001). However, I should note that those previous studies focused on LSA's performance associated with other linguistic tasks such as information retrieval (Dumais) or document categorization (Nakov et al.), rather than document relatedness judgments.

In sum, the main purpose of Chapter 3 was to find the optimal parameters of LSA, yielding the highest agreement between humans and LSA. With the help of numerous background documents, the highest correlation obtained between humans and LSA was 0.70, which was significantly greater than that of Pincombe (r = 0.60, 2004). In the next chapter, using the optimal parameters of LSA reported here, I will identify document pairs whose

relatedness is judged radically differently between humans and LSA (i.e., the misjudged document pairs).

Chapter 4. Identifying the Misjudged Document Pairs

The purpose of Chapter 4 is to identify extremely misjudged document pairs by comparing human document relatedness ratings and LSA-cosine scores based on the optimal parameters determined in the previous chapter. Lee et al. (2005) hypothesized that the main source of the mismatch is document pairs whose LSA-cosine scores are not as high as human ratings (i.e., misses). According to this hypothesis, one can expect that the misjudged document pairs would typically have relatively high human ratings with relatively low LSA-cosine scores.

To compare human ratings and LSA-cosine scores, I normalized the 1,225 human ratings (*z*-score normalization) obtained in Pincombe (2004). I also normalized the corresponding 1,225 LSA-cosine scores obtained based on the optimal parameters. The right panel of Figure 5 shows the correspondence between humans and LSA, where each circle represents a document pair.

According to Lee et al. (2005) the misjudged document pairs should be primarily observed on the upper left side of the scatter plot. To test this hypothesis and, more importantly, to identify the misjudged document pairs, I compared the normalized ratings of human and LSA. Specifically, document pairs with absolute *z*-score differences of 1.0 or greater (i.e., |human *z*-score – LSA *z*-score| $\geq$ 1.0) were selected as candidates of the misjudged document pairs. There were 203 such document pairs (17% of the 1,225 document pairs), which are marked in red in Figure 6.

Seemingly consistent with the hypothesis of Lee et al. (2005), there was a group of misjudged document pairs (red circles) observed in the upper left part of the scatter plot (Figure 6). Normalized LSA-cosine scores of those pairs were not as high as the normalized human ratings, with minimum difference of 1.0. Hereafter, this type of misjudged document pairs will be called the *missed document pairs*, regarding the human responses as correct judgments.

Notably, however, there was another group of misjudged document pairs on the opposite side of the scatter plot. Normalized LSA-cosine scores of those pairs were not as *low* as the normalized human ratings, with a minimum difference of 1.0. Such document pairs were unexpected, based on the hypothesis of Lee et al. (2005). I will discuss the potential reasons for observing this unexpected mismatching pattern in the Discussion section of this chapter. Hereafter, this type of misjudged document pairs will be called the *false positive document pairs*.

Out of the 203 candidates of misjudged document pairs, there were 120 (59%) misses and 83 (41%) false positives. Among the 120 candidates of the missed document pairs, those with relatively high human raw-score ratings, operationally defined as > 2.5 (*z*-score > 1.23), were regarded as the best candidates to be misses[5]. Similarly, among the 83 candidates of the false positive document pairs, those with relatively low human raw-score ratings, operationally defined as < 1.6 (*z*-score < -0.05), were regarded as the best candidates to be false positives. Based on these criteria, I selected 10 candidates of misses (average *z*-score difference of 2.36), and 12 candidates of false positives (average *z*-score difference of 1.57), out of the original 203 misjudged document pairs. I selected the 22 candidates so that no document was used in more than one pair within each misjudgment category (i.e., miss or false positive), to cover as many different documents as possible[6]. There were 34 individual documents in the 22 candidates of misjudged document pairs, which are marked with black dots inside of the red circles in Figure 6.

Human Experiment 1: Replication of the Previous Human Relatedness Ratings

[5] One could select document pairs with relatively low human raw-score ratings and further lower LSA-cosine scores as missed document pairs. However, document pairs with true relatedness, which is operationally defined as high human ratings (e.g., > 2.5) offer better examples of misses.

[6] The misjudged document pairs will be further studied in the subsequent chapters to better understand human and LSA's document relatedness judgments. If the misjudged document pairs have only a few documents in them, for example, if there are only eight different documents in the 22 misjudged document pairs, then the findings of the later chapters could be merely a document-specific phenomenon rather than revealing a general phenomenon of human and LSA's document relatedness judgments.

The 22 candidate pairs identified above are valid misjudged document pairs only if the human ratings used in the selection process validly reflect human judgments associated with those pairs. In Pincombe (2004), from which the original human ratings were obtained, each participant rated more than 100 document pairs, and only about 10 ratings (± 2) were obtained for each document pair. Because of this relatively small sample of ratings, it is possible that the high average ratings in the current miss candidates (> 2.5) were obtained by chance, which would disqualify them as a missed document pair. Likewise, the low average ratings in the current false positive candidates (< 1.6) might have been obtained by chance. It might seem unlikely that several of them were badly misjudged by the human samples; however, note that the candidate misses and false positives were identified by selecting pairs with the most extreme misjudgment from the original sample of 1,225 document pairs. Given the likelihood that some luck was involved in producing the extreme values, one would naturally expect that a replication would reveal considerable regression back to the true population mean. Therefore, to ensure that the human ratings obtained by Pincombe validly reflect the true judgments of humans, the 22 document pairs were rated again by another group of human readers.

Method

Participants. Twelve undergraduate students from the University of New Mexico participated in exchange for course credit.

Stimuli. Out of the 1,225 document pairs used in Pincombe (2004), 203 document pairs with an absolute *z*-score difference of 1 or greater between humans and LSA were selected. Among them, 10 pairs with high human ratings (> 2.5), and 12 pairs with low human ratings (< 1.6) were selected as candidates of miss and false positive document pairs, respectively. They were also selected so that, within each misjudgment category (miss or false positive), no

document was used in more than one pair (there were 34 unique documents). However, nine documents from one category were also used in the other category. Such overlap was inevitable because there were only 50 documents to begin with. The two documents of each of the 22 misjudged document pairs were printed side-by-side on a piece of paper.

Procedure. In the beginning of the experiment, participants read and signed on the consent from. Participants self-reported whether they were native English speakers or not. Then, they were asked to rate, on a scale of 1 to 5 (1 indicated "not related at all," and 5 indicated "strongly related"), how similar they felt the given documents were to each other. As in Pincombe (2004), participants were not given any instructions as to the strategy they should use to make their judgments. They rated the relatedness of the 22 document pairs in a fixed, pseudo-random order.

Results

Two of the participants were not native English speakers, so their data were excluded from the remaining analysis. For each document pair, scores from the 10 remaining participants were obtained and averaged. Table 1 shows the average ratings obtained in the current study as well as in Pincombe (2004).

The correlation between the two studies across the 22 pairs was 0.71. Participants in the current study generally rated the false positive pairs higher than in the previous study (Pincombe, 2004). Such an increase could be due to the different context in which participants rated the pairs: they rated only 22 document pairs in a fixed order in the current study, rather than rating more than 100 document pairs in random orders as in Pincombe. However this seems a trivial change. Another explanation (noted above) is that, due to the way the 22 documents were selected (the most extreme misjudged document pairs out of 1,225 pairs), there was some

regression back to the mean. Consistent with this regression-to-the-mean explanation, for the missed document pairs the new data generally showed lower ratings (average of 2.86) than Pincombe (average of 3.15). Also, for the false positive document pairs the new data generally showed higher ratings (average of 1.82) than Pincombe (average of 1.18). Among the 10 candidates of missed document pairs, six pairs still had relatively high human raw-score ratings, equal to or greater than 2.80. These documents were selected as the final missed document pairs for further study. Among the 12 candidates of the false positive document pairs, six pairs had relatively low human raw-score ratings, equal to or lower than 1.6. They were selected as the final false positive document pairs. Table 2 shows examples of the missed and false positive document pairs.

Discussion

Although Lee et al. (2005) hypothesized that misses would be the main type of misjudgment by LSA, the current analyses showed that there were nearly as many other document pairs that caused the opposite misjudgment type, false positive. However, there were still significantly more missed document pairs than false positives (59% vs. 41% of misjudged pairs) [$\chi^2(1) = 6.74$, $p = .01$]. Given this misjudgment, the next question is what could have caused the misses and false positives by LSA? To answer this question, I first empirically verified the assumption that LSA's document relatedness is primarily based on the relatedness between words in the document pairs. Then, I generated hypotheses regarding the potential causes of the misjudgments by LSA.

If LSA bases its document relatedness judgments largely on the degree of word relatedness in document pairs, there should be a high correlation between LSA-cosine scores for document relatedness and the average LSA-cosine scores for the corresponding word pairs

within document pairs. To test this hypothesis, I sampled every possible word pair from each of the 1,225 document pairs, ignoring stopwords. Then I obtained LSA-cosine scores of the word pairs in each document pair based on the optimal parameters determined in the previous chapter, and averaged them to produce an overall word-relatedness score of each document pair. The correlation between LSA's document relatedness scores and the averaged word relatedness scores was 0.73 ($p < .01$). This result shows that LSA's document relatedness can be explained well by the word relatedness. This also means that LSA can be used to emulate, with a reasonable amount of accuracy, the contrast model. However, I should also note that the imperfect correlation suggests that LSA is not judging document relatedness entirely based on word relatedness.

Given the above demonstration that LSA bases document relatedness judgment largely on word relatedness, a potential reason that LSA misses document relatedness is that LSA misses the relatedness of some word pairs in documents. In fact, there are various reasons that LSA can miss word relatedness. For example, although "United States", "US", "U.S.", and "U.S.A" refer to the same country, these words may be excluded from the word-by-document matrix during the pre-processing due to the fact that they have two words, to their short length, or to the use of special characters, whereas humans presumably utilize them to judge the relatedness of documents including those words. Therefore, LSA may judge document pairs including such words to be less related than humans would (i.e., leading to a miss). Also, there are some words in the target documents that were not included in LSA's background documents (or included very rarely), preventing LSA from correctly judging the relatedness of documents including those words.

Similarly, a potential cause of false positives by LSA is that LSA mistakenly perceives relatedness between words that are in fact unrelated. Depending on the contents of background documents, LSA may recognize some word pairs to be more related than humans would and, as a result, judges document pairs including those word pairs to be more related than humans do. The above hypotheses regarding misses and false positives of LSA are based on an assumption that LSA's word relatedness judgments are often flawed.

An alternative hypothesis regarding misses and false positives of LSA is that LSA reliably captures the overall word relatedness in document pairs (as it is often hypothesized to do so) and judges the document relatedness accordingly, whereas human document relatedness ratings relies heavily on many other factors. If human document relatedness does not closely correspond to the degree of word relatedness in documents (i.e., contradicts the contrast model), then LSA's judgments should be quite different from human judgments. I should note, however, that this hypothesis does not necessarily assume that the contrast model cannot explain human document relatedness judgments at all. Their judgments can still be dependent on the degree of word relatedness to a certain degree.

To what degree can human document relatedness be explained by human word relatedness judgments? To answer this question, one requires a data set that includes both human relatedness scores of document pairs and human relatedness scores of every possible word pair within the corresponding document pairs (or even the majority). Therefore, in the next chapter, I collected human word relatedness scores from the misjudged document pairs identified in this chapter.

In sum, in the current chapter I identified the two types of misjudgment by LSA despite optimal parameters: misses and false positives. Regarding the potential causes of the two

misjudgment types of LSA, a natural hypothesis is that LSA misperceives word relatedness: LSA misses some word relatedness from the missed document pairs, and falsely detects word relatedness from the false positive document pairs. There is also an alternative hypothesis: human document relatedness is not largely dependent on the degree of overall word relatedness (as the contrast model suggests and as LSA does). To empirically test these hypotheses, human word relatedness rating data is critical, and thus such data were collected, as reported in the next chapter.

Chapter 5. Applying the Contrast Model to Human Document Relatedness Judgments

This chapter directly examines whether the contrast model can explain human document relatedness judgments in terms of word relatedness. Because LSA's document relatedness judgments are also based largely on the word relatedness, the answer will also help clarify the most likely sources of misjudgment of LSA. To accomplish this goal, I need a data set that includes both human document relatedness scores and human word relatedness scores of every possible word pair within the document pairs. Because there is no such complete, pre-existing data set, in Chapter 5, I conducted a word relatedness judgment experiment based on the 12 misjudged document pairs identified in Chapter 4.

Specifically, I sampled every possible word pair from the 12 misjudged document pairs, ignoring stopwords, and collected word relatedness scores from human readers. Participants rated the relatedness of those word pairs without knowing which documents they came from. According to the contrast model, missed document pairs with high human document relatedness ratings (average raw-score rating of 3.30) should show a high degree of word relatedness, while the false positive document pairs with low human ratings (average raw-score rating of 1.35) should show a low degree of word relatedness. Alternatively, human document relatedness may not closely correspond to the degree of word relatedness; that is, humans might mainly rely on other indicators of document relatedness, such as higher-order structural properties (Gentner, 1983), or focus only the keywords. In this case, one can expect a weak relationship between human document relatedness ratings and human word relatedness ratings.

Human Experiment 2: Word Relatedness Judgment

Method

Participants. Four-hundred and twenty-one undergraduate students from the University of New Mexico participated in exchange for course credit.

Stimuli. From the 12 misjudged document pairs, I found 12,023 word pairs after removing stop words, and 23 proper nouns participants would not know, such as Ceja (a codename of a corruption task force team used by the Australian police). Each word was presented in its original form in the document with a few exceptions[7]. After I removed 23 obviously unrelated word pairs to produce a round number that would divide evenly into sets, 12,000 word pairs remained. Those 12,000 word pairs were separated into 40 sets with 300 word pairs each (40 X 300 = 12,000). Word pairs were presented in a random order one-by-one on a computer screen below a 5-point rating scale as shown in Figure 7.

Procedure. The current study was an on-line experiment. Participants signed up for the study through the University of New Mexico psychology department's web advertising system. They read an on-line consent form, and clicked a message box saying 'I agree to participate in this study' at the bottom of the consent form, to participate in the study. In the next page, they were instructed to give lower numbers for less related word pairs and higher numbers for more related pairs using the 5-point scale. If they did not know any given word or the relationship of the given words, they were instructed to choose "Don't Know" button, located on the left side of the scale. After making a response, by hitting the "Next" button, participants could move to the next trial. Each participant made 315 responses where the last 15 trials were repeated word pairs randomly selected from the previous 300 trials. These repeated trials were used to measure each participant's response reliability, which is the correlation (*r*) between the two sets of 15 trials.

[7] If a document includes two or more words from the same word root, such as Australia and Australian, then the derivatives were presented together using " / ", for example, 'Australia/Australian'. There was another exception in which a news article (Article 18) included 7 African country names. Those names were presented together separated by " / ". Other than these cases, words were presented in the original form, as shown in Figure 7.

Results

For each of the 40 word-pair sets, there were at least 10 participants (except for the last two sets with 9 participants). To ensure data quality, I excluded 49 out of 421 (12%) participants' data based on the following criteria: (a) participants took less than 10 minutes to complete the task (6 participants, 1%), (b) the number of "Don't Know" responses was greater than 50 (5 participants, 1%), (c) reliability correlations were not positive (33 participants, 8%), and (d) agreement[8] were not positive (5 participants, 1%).

Table 3 shows the average relatedness ratings of all word pairs in each document pair (see column 4). The critical finding is that the average word relatedness rating of 3.00 in the six missed document pairs was not significantly different from the average word relatedness rating in the six false positive document pairs [2.85; $t(10) = 1.85$, $p = .095$]. Note that, despite nearly equivalent word-relatedness scores, missed and false positive document pairs received radically different relatedness ratings from human readers: an average raw-score rating of 3.30 and 1.35, respectively, for the miss and false positive pairs (based on the experiment reported in Chapter 4). Thus, human word relatedness ratings failed to capture the differences in human document relatedness ratings between the two sets of misjudged document pairs.

Discussion

According to the contrast model, the average relatedness of the possible word pairs within a document pair should predict the overall relatedness of that document pair. However, the word relatedness data obtained from human readers failed to capture the radical relatedness differences between the two sets of misjudged document pairs. The current result suggests that the contrast model needs some modification to satisfactorily explain human document

[8] Agreement was the average of the correlations between one participant and all the other participants assigned to the same word-pair set.

relatedness judgments. The results also help explain why LSA, which emphasizes word relatedness, often misjudges the relatedness of document pairs.

As Gentner (1983) has pointed out the contrast model may be too simple to explain complex semantic judgments by humans, because humans may use higher level information in documents besides mere word overlap. Therefore, humans may judge some document pairs to be highly related despite relatively low word relatedness (i.e., misses). On the other hand, the other case, the false positives (high word relatedness yet low document relatedness) is particularly puzzling from the perspective of the contrast model. The low human ratings of the false positive document pairs indicate that humans ignored the relatively high word relatedness and judged them to be unrelated based on some other factors. What exactly are these factors? Below, to answer this question, I further examined the false positive document pairs.

Qualitative Analysis of the False Positive Document Pairs

Table 2 (lower panel) shows a concrete example of a false positive document pair. The average human word relatedness rating of this document pair was 3.10 (see Table 3) on a scale of 1 to 5, which was almost the same as, yet slightly *greater* than the averaged human word relatedness ratings in the six missed document pairs (3.00). Based on the document pair's overall word relatedness, the contrast model should predict that this document pair would be judged to be related by both LSA and humans. Consistent with this prediction, LSA assigned the pair a relatively high cosine rating (0.18, $z = 0.8$). However, humans judged the pair to be unrelated by assigning an average rating of 1.10 ($z = -0.76$) on the scale of 1 to 5.

As a first step of examining this specific false positive document pairs, I examined word pairs that were judged highly related by both humans and LSA in this false positive document pair. Table 4 shows the 30 most highly related word pairs (out of 798 word pairs) judged by

humans and LSA in this false positive document pair. Visual inspection of Table 4 suggests that some of LSA's highly related words did not make as much sense as the highly related words defined by humans. For example, from Table 4, we can find that LSA judged word pairs that include Australia, such as "Australia ~ Target" ($z = 2.32$), to be highly related. Even more important, some words in LSA's highly related word pairs do not appear to be important (i.e., keywords) in representing each document's topic. For example, the word pair "document ~ design" which was judged to be most related by LSA may not contribute much in characterizing the meaning of each document. Still, LSA may have judged the false positive document pair to be related based on the highly related word pairs, regardless of the keyness of the words. On the other hand, for humans, the relatedness of non-keywords may not affect their judgments of this specific document pair's relatedness.

Based on the above observation, I modified the contrast model to better explain human document relatedness judgments. This modified model hypothesizes that people use word relatedness in judging the relatedness of document pairs (as assumed in the contrast model), but words that are key to the documents' meaning are given more weight. That is, word relatedness is especially indicative of document relatedness when the words are key. According to this modified model, document pairs with many highly related words could still be regarded as unrelated by humans if those words are not keywords. However, computer models may not be fully utilizing the keyness of the related words in making their document relatedness judgments. Therefore, they tend to make false positive judgments with such document pairs. Although this modification of the contrast model was primarily inspired by the examination of the false positive document pairs, it may also be used to explain how humans judge the relatedness of the missed document pairs differently from LSA. Specifically, the missed document pairs could

have a low degree of overall word relatedness with high degree of keyword relatedness. Based on the high keyword relatedness, humans could judge the pair to be more related than LSA does. I will call this modified version of the contrast model the key-features overlap model, which will be empirically tested in the next chapter.

Chapter 6. Test of the Key-Features Overlap Model

In Chapter 5, I showed that the contrast model failed to explain human document relatedness judgments. Specifically, it could not readily explain why humans judged certain document pairs (the false positives) with high word relatedness to be unrelated. To better explain human document relatedness judgments, I proposed a modified version of the contrast model, the key-features overlap model, which argues that human document relatedness judgment is primarily based on the relatedness of *keywords* in document pairs, while computational models including LSA do not adequately utilize the keyness of words.

The validity of this key-features overlap model critically depends on to what degree the keyness of words plays a role in judging document relatedness. Therefore, to test this model, I conducted a human experiment in which participants identified the keywords in each of the 50 target news articles. According to the key-features overlap model, the relatedness among keywords should predict human document relatedness better than the relatedness of all word pairs in a document pair. Also, if LSA does not use the keyness of words in judging document relatedness, then LSA-cosine scores between keywords are not necessarily good at predicting LSA's document relatedness comparing to LSA's overall word relatedness scores.

Human Experiment 3: Identifying the Keywords in Documents

In the current experiment, participants read individual news articles one-by-one, and identified 10 keywords from each article. They were also instructed to rank-order the 10 selected keywords by their importance in representing each article's topic.

Method

Participants. One-hundred undergraduate students from the University of New Mexico participated in exchange for course credit.

Stimuli and procedure. The 50 news articles were separated into 3 groups with 14, 15, and 21 documents each. In the beginning of the experiment, participants self-reported whether they were native English speakers. Each news article was presented on the computer screen one-by-one (see Figure 8). Participants were instructed to read the whole article first without selecting any keywords. To encourage the initial reading of the whole document, the mouse cursor was hidden for 30 seconds after document presentation. Then, a pop-up message saying "OK to begin" appeared, the mouse cursor became visible, and participants could start keyword selection. Left-clicking on a word highlighted the word, and made the counter located on the left side of the screen increase by 1. Right-clicking on a previously highlighted word unhighlighted the word, decreasing the counter by 1. Stop words were not clickable. Only after participants selected 10 keywords, could they press the "10 selected keywords" button, which showed the 10 selected words in a listbox. If participants hit this button without selecting 10 keywords, then an error message instructed them to select 10 keywords. After they reviewed their own 10 selected keywords, they hit the "Start Ordering" button. Then, the same word list appeared in a separate listbox where participants could move each word's position. Participants were instructed to place the most important keyword at the highest position of the listbox, and place the second most important word at the second-highest position, and so on. After ordering the 10 keywords, participants pressed the "Save and Proceed" button to move on to the next trial.

Keyness calculation and keyword selection. To determine each word's keyness score, I used the following weighting scheme. Each word's frequency in each rank position (i.e., 1^{st} ~ 10^{th}) was multiplied by the inverse of its rank (i.e., 11 minus its rank), and then these products were summed. These sums were then divided by the number of participants in the word's document set to compensate for the effect of different numbers of participants across document

sets. For example, if a word "Balance" was observed five times in the first rank, and four times in the second rank and not observed in other ranks until it was found once in the 10th rank, with 30 participants, then its keyness score was obtained as follows:

Keyness of "Balance" = [5 X (11 – 1) + 4 X (11 – 2) + 0 X (11 – 3) + … + 0 X (11 – 9) + 1 X (11 – 10)] / 30 = (50 + 36 + 0 + … + 0 +1) / 30 = 87 / 30 = 2.90.

The top 30% of the words within a document (but not stopwords) were selected as keywords based on their keyness scores. The agreement among participants with respect to their keyword selections was estimated using the Jaccard similarity coefficient, which is defined as the size of the intersection divided by the size of the union of the features from two objects,

$$\text{Jaccard similarity coefficient} = \frac{size\ of\ (A \cap B)}{size\ of\ (A \cup B)}$$

Results

There were 24, 33, and 25 native English speaker participants, judging the first, second and third document set, respectively. Data from non-native English speakers were excluded from further data analysis. The minimum, maximum, average, and standard deviation of the keyness scores were 0.03, 21.40, 1.93, and 1.98, respectively. The average Jaccard coefficient across the 50 documents was 0.30, suggesting a moderate level of agreement among participants with respect to their keyword selections.

The main purpose of this chapter is to test whether the relatedness of keywords in a document pair can better predict the relatedness of the document pair, compared to using the

relatedness of all words. To perform this test, I used the human word relatedness scores obtained from the 12 misjudged document pairs in Chapter 5. Table 3 (see 5th column) shows the average of keyword relatedness scores in each misjudged document pair. The average relatedness score among keywords was 3.13 in the six missed document pairs (i.e., those with high human document relatedness ratings), which was significantly greater than the average relatedness score among keywords in the six false positive document pairs (i.e., those with low human document relatedness ratings) [2.64, $t(10) = 3.02$, $p = 0.01$]. Recall that, in Chapter 5, using the average word relatedness scores across all words, this difference between the two misjudgment types was not significant (average word relatedness score of 3.00 and 2.85, in the missed and false positive document pairs, respectively). Also the mean difference in keyword relatedness scores between the two misjudgment types was three times as great as the mean difference of overall word relatedness scores in the two misjudgment types, showing marginally significant interaction effect [0.49 vs. 0.15; $F(1, 20) = 3.61$, $p = .072$]. This result suggests that keyword relatedness can capture at least some of the radical differences in document relatedness between the two misjudgment types. In other words, by taking keyness into account, it should be possible to fix (at least partially) LSA's most extreme misses and/or false positives.

I also examined whether LSA's keyword relatedness scores can predict human document relatedness ratings better than its overall word relatedness scores. The correlation between LSA's overall word relatedness scores and human document relatedness scores was 0.60. However, this correlation increased to 0.70 when the LSA word relatedness scores were obtained from only keywords ($z = 4.3$, $p < .01$). That is, using the LSA-cosine scores from only 30% of words (i.e., keywords) in each document, one can predict the corresponding human document relatedness better than using 100% of the words. Also, estimating human document relatedness

based on LSA-cosine scores of keywords is as accurate as that of the typical LSA procedure that yielded a correlation of 0.7 with human ratings (see Chapter 3). The current result shows that utilizing keywords can be an efficient and effective approach to predicting human document relatedness. Also, it supports the hypothesis that when humans use the relatedness of words to judge document relatedness, they primarily focus on the relatedness of keywords rather than all words.

To demonstrate the usefulness of utilizing keywords within LSA to judge document relatedness, we can repeat the same analysis (predicting human document relatedness based on LSA-cosine scores of word relatedness) but now with a randomly chosen set of 30% of the words (which may also include keywords) in each document instead of a set of keywords. The average correlation between the averaged LSA-cosine scores of randomly chosen 30% of words and human document relatedness ratings was only 0.5. That is, we can explain an additional 24% (49% – 25%) of the variance in human document relatedness judgments by using the cosine scores between keywords rather than those of randomly selected words.

Discussion

In the previous chapter, I addressed the potential strengths of humans and weaknesses of LSA with respect to their relatedness judgments of false positive document pairs. Based on a qualitative examination of the false positive document pairs, I hypothesized that humans base their document relatedness judgments on a subset of keywords, ignoring the relatedness between non-keywords, whereas LSA bases its document relatedness judgments on the degree of overall word relatedness, regardless of their keyness (the key-features overlap model).

To test the key-features overlap model, in this chapter, I conducted a human experiment in which participants identified the keywords of each document. Based on the human data, I

determined the keywords that best represented the meaning of each document (top 30%), and examined whether the degree of relatedness between keywords can predict human document relatedness judgments. Results showed that the keyword relatedness can capture the differences in misjudged document pairs [$t(10) = 3.02$, $p = 0.01$]. The current result supports the hypothesis that people base their judgments of document relatedness not on overall word relatedness but rather on the relatedness of words that are key to defining a document's meaning.

I also showed that LSA was able to predict human document relatedness based on only a relatively small number of keywords as accurately as when it is based on all the words. This finding supports the hypothesis that human document relatedness relies primarily on the relatedness among keywords (i.e., the key-feature overlap model). To generalize the current study, replications based on other document sets (that have human document and word relatedness rating data) are required.

Modifying LSA to Utilize Keyness

To examine to what degree LSA happens to utilize the keyness of words in making its document relatedness judgments, I investigated how well LSA's document relatedness correlates with LSA's overall word relatedness versus LSA's keyword relatedness. The former correlation was 0.73 (as reported in Chapter 4), which was numerically smaller than the latter correlation of 0.78, although the difference was only marginally significant[9] ($z = 1.79$, $p = 0.07$). This small difference suggests that LSA does not fully utilize the relatedness between keywords in determining the relatedness between documents.

[9] This trend may arise because keywords are often used more than once in a document. LSA's document relatedness, which is a reflection of word overlap between documents, is already sensitive to word frequency. Therefore, LSA document relatedness may correlate better with keyword relatedness than overall word relatedness.

Along with the above observation, the key-features overlap model suggests that LSA's document relatedness judgments can be improved by modifying it so that it further utilizes the relatedness among keywords. Specifically, it suggests that document pairs with relatively high overall word relatedness and low keyword relatedness should be the strong candidates for false positive judgments. Similarly, document pairs with low overall word relatedness and high keyword relatedness should be the strong candidates for miss judgments. To reduce both misses and false positives, according to the key-features overlap model, it is critical for LSA to accurately judge the relatedness of keyword pairs. Then, how can we improve LSA's keyword-relatedness-detection ability?

Among various approaches proposed to enhance computational algorithms' document relatedness judgments, some have attempted to utilize the keyness of words in documents. For example, a group of researchers attempted to judge the relatedness of two documents written in two different languages by considering the relatedness of keywords in the two documents (Steinberger, Pouliquen, & Hagman, 2002). Other researchers showed that using only 30~60% of keywords from each document can improve a computational algorithm's document classification (Kang, 2003). While these approaches focused on directly extracting keywords of a document and calculating their relatedness to estimate document relatedness, approaches introduced below inspired by the key-features overlap model focus on how to modify the target and background documents subject to LSA so that the keywords in target documents are emphasized and so that LSA can judge the relatedness of those keywords better based on more keyword relevant background documents.

As demonstrated in Chapter 3, as the number of background documents increases, LSA's document relatedness judgments improved. To not only increase the numbers of background

documents, but also increase LSA's keyword-relatedness-detection ability, I propose to use the keywords' dictionary information as background documents. For example, Wikipedia articles selected by keywords as a search query can increase the size of background documents and may also increase LSA's keyword-relatedness-detection ability. In addition to this, one may repeat keywords of each document so that LSA can easily capture the keywords of the document. In the next chapter, these modifications inspired by the key-features overlap model will be tested.

Chapter 7. Modification of Documents Subject to LSA Based on the Key-features Overlap Model

The key-features overlap model suggests that if LSA were to utilize the keyness of words in judging document relatedness, it would produce more human-like judgments. In the current chapter, I report attempts to modify the target and background documents subject to LSA so that LSA can capture the relatedness of keywords better. Specifically, I attempted the following three approaches: (1) repeating keywords of each document as a proportion of their keyness (the keyword-repetition condition), (2) providing Wikipedia articles retrieved based on keywords as a search query (the Wikipedia-background condition), (3) the combination of these two methods (the combination condition). These approaches were intended to emphasize the keywords of documents (keyword-repetition condition) and encourage LSA to capture the relatedness of keywords better based on direct knowledge about the keywords (the Wikipedia-background condition).

In the Wikipedia-background condition, among the 1,471 keywords of the 50 target news articles defined in Chapter 6, I chose 409 keywords and retrieved a corresponding Wikipedia article. The selected keywords were mostly noun keywords, because nouns tend to have fewer possible meanings and usages than verbs. That is, two documents sharing the same noun are more likely to be related than ones sharing the same verb. I also chose keywords that are proper nouns, because they may have relatively low frequencies and be ignored during the pre-processing due to their low frequency. By including Wikipedia articles as background documents that were retrieved based on the proper noun keywords, corresponding keywords should survive the pre-processing process and be utilized in judging document relatedness by LSA.

Method

Stimuli. The 50 target news articles and 4,486 background documents used in Chapter 3 were employed as the main stimuli. In the Wikipedia-background condition, Wikipedia articles retrieved based on the 409 keywords as a search query were selected as additional background documents. All contents contained in each Wikipedia article comprised a single background document. However, I excluded some words commonly occurring in Wikipedia articles, such as Wiki, article, link, and so on.

Procedure. In the keyword-repetition condition, the keywords of each document were repeated as a function of their keyness, so that keywords with higher keyness occurred more frequently than keywords with lower keyness. Specifically, keywords were stratified into 10 classes based on their keyness determined in Chapter 6, and keywords in the top 10% were repeated 10 times, and ones with the importance of the next top 10% were repeated 9 times, and so on, until the keywords with the lowest keyness were repeated just once. For example, the first news article has 30 keywords and the first 3 keywords with the highest keyness were repeated 10 times, and the next 3 keywords with the second highest keyness were repeated 9 times, and so on.

In the Wikipedia-background condition, the 409 Wikipedia articles were added to the corpus of 4,486 background documents. Finally, in the combination condition, the target documents with keyword-repetition and the 4,486 background documents along with the 409 Wikipedia-background documents were subjected to LSA.

For each condition, I tested the same combinations of parameters (weighting schemes and dimensions) used in Chapter 3, to find the maximum correlation between humans and LSA across conditions. Specifically, I tested every possible combination of the three local weightings (term frequency, logarithmic, and alternative logarithmic) and three global weightings (idf,

entropy, and p-inverse). I also tested every combination of these weighting schemes at each of the 23 dimensions used in Chapter 3 (10, 20, 30, 40, 50, 100, 150, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 350, 360, and 364).

Results

The correlation between humans and LSA in the three conditions (the keyword-repetition, the Wikipedia-background, and the combination condition) across the various combinations of weighting schemes and dimensions behaved similarly as in Chapter 3. That is, the local and global weighting schemes had much less effect than the dimensionality, and the maximum correlation was obtained in the middle range of the dimensions, between 100 and 200. Figure 9 shows the highest correlation between humans and LSA in the three conditions, as well as the results from Chapter 3 (r = 0.7), where the original 50 target and 4,486 background news articles were subject to LSA.

The maximum correlation of 0.71 in the keyword-repetition condition was obtained from the term frequency local – entropy global weighting at the dimensionality of 200. The maximum correlation of 0.72 in the Wikipedia-background condition was obtained from the alternative logarithmic local – idf global weighting at the dimensionality of 220. Finally, the maximum correlation of 0.73 in the combination condition was obtained from the logarithmic local – p-inverse global weighting at the dimensionality of 150. The dashed line in Figure 9 shows the highest correlation of 0.72 between humans and LSA obtained in a previous study that used the same target news articles (Gabrilovich & Markovitch, 2007).

Discussion

In this chapter, following the key-features overlap model, three modifications of the target and background documents were attempted to emphasize the keywords in target

documents and provide LSA with better keyword-relevant background documents. In the combination condition that involved both approaches, the overall correlation between human ratings and LSA-cosine scores across the 1,225 document pairs showed a correlation of 0.73 which is marginally greater than the correlation of 0.7 obtained in Chapter 3 without any modifications ($z = 1.52$, $p = 0.06$, one-tailed).

The impact such modification was even more impressive when examining the degree of misjudgments by LSA in missed and false positive document pairs. Before the modifications, the averaged *z* score differences between humans and LSA for missed and false positive document pairs were 1.57 and 1.42, respectively. In the combination condition, those two values decreased to 1.25 [$t(208) = 3.91$, $p < .001$] and 1.03 [$t(128) = 4.62$, $p < .001$], respectively, suggesting that the modification reduces the degree of misjudgment by LSA in those two types of misjudgments.

Gabrilovich and Markovitch (2007) who obtained the highest agreement between humans and LSA so far ($r = 0.72$) used 241,393 Wikipedia articles as background documents. In the current study, for background documents, I used 4,486 news articles that came from the same source as the target documents and 409 Wikipedia articles that derived by keywords as a search query. That is, based on only about 20% (4,895 documents) of the background documents employed in the previous study, LSA showed a similar level of agreement. The current result suggests that a modification of the source documents subject to LSA following the key-features overlap model can be promising for improving LSA's document relatedness judgments.

In this chapter, the modification was made based on the keywords selected by human readers (in Chapter 6). However, in practice, there would be no such keyword data generated by humans. Therefore, to apply the proposed modification to real-world data, one may need to

determine keywords of each document using keyword-extraction algorithms such as that proposed by Matsuo and Ishizuka (2003).

Chapter 8. General Discussion

The current study had three overarching goals: a) to better understand how humans judge document relatedness, b) to modify Tversky's (1977) contrast model to better explain human document relatedness judgments, and c) to apply the modified model to improve LSA's document relatedness judgments. To accomplish these goals, I discussed relevant theories, computational techniques, conducted several human experiments, and modified source documents subject to LSA.

In Chapter 2, I introduced psychological models of human similarity judgments including the contrast model, and also introduced computational models of document relatedness judgments such as the vector-space model and its variant LSA. The underlying mechanisms of document relatedness judgments assumed by the contrast model and the vector-space model were viewed to be similar to each other. In Chapter 2, I also introduced a previous study that examined the degree of agreement between humans and LSA with respect to their document relatedness judgments (r = 0.6; Pincombe, 2004). Regarding the mismatches between humans and LSA, Lee and colleagues (2005) hypothesized that document pairs that are judged to be related by humans but not by LSA (i.e., a "miss" by LSA) would be the main source of mismatch. To test their hypothesis under LSA's optimal parameter setting, in Chapter 3, I tested various parameters of LSA. The highest correlation was 0.7, obtained with a large number of background documents (the 4,486-document background condition). Comparisons of human and LSA's document relatedness ratings showed that there were missed document pairs, as Lee and colleagues hypothesized. An unexpected result, given Lee and colleagues' hypothesis, was a large number of false positive document pairs, which were judged to be related by LSA but not by humans.

A potential cause of misses by LSA is that LSA misses relatedness between words. Similarly, a potential cause of false positives by LSA is that LSA mistakenly perceives relatedness between words that are in fact unrelated. An alternative hypothesis regarding misses and false positives of LSA is that human document relatedness ratings do not closely correspond to the degree of word relatedness in documents (i.e., contradicts the contrast model). Then LSA's judgments should be quite different from human judgments. In Chapter 5, to directly test the contrast model, I collected human word relatedness ratings from the misjudged document pairs. The results showed that overall word relatedness in document pairs could not capture the radical differences of human document relatedness ratings between the missed and false positive document pairs. Importantly, comparison of the highly related word pairs determined by humans and LSA (see the last part of the discussion of Chapter 5) indicated that, to judge document relatedness, humans primarily focus on the relatedness between keywords while LSA relies on the overall word relatedness without considering each word's keyness. This modified view of the contrast model was called the key-features overlap model. In Chapter 6, I reported a human experiment of keyword identification that was conducted to test the key-features overlap model. The results showed that the relatedness of keywords is a better predictor of the relatedness of document pairs than the relatedness of all words, supporting the key-features overlap model. The key-features overlap model suggests modification of LSA so that it can utilize the relatedness among keywords (e.g., using keyword-relevant Wikipedia articles as background documents). In Chapter 7, I tested such modifications of documents subject to LSA according to the key-features overlap model. Specifically, I repeated keywords in each document or/and provided keyword-relevant Wikipedia articles as background documents. Results showed that these attempts tended to enhance LSA's document relatedness judgments and reduce the severity

of misses and false positives. These findings demonstrate the usefulness of modifying LSA following the key-features overlap model.

Miss and False Positive vs. Precision and Recall

In computer science, the typical measures used for evaluating an algorithm's performance, especially its information-retrieval performance, are precision and recall (Powers, 2011). *Precision* is the number of retrieved documents that are relevant to a given query over the total number of relevant documents in a given database. *Recall* is the number of relevant documents retrieved over the total number of documents in the database. Suppose that there are 10 documents in a given database with 5 relevant documents and 5 irrelevant documents to a given query. If a computer algorithm retrieved 5 documents, in which 3 of them were relevant and 2 were not, then the algorithm's precision is 3/5, while its recall is 3/10. The two irrelevant documents retrieved correspond to false positives, while the two relevant documents that were not retrieved correspond to misses. Precision and recall measure an algorithm's accuracy and sensitivity, respectively, and can be regarded as indices of the level of success. On the other hand, the misses and false positives reflect errors of an algorithm. In the current study, given that LSA already shows a relatively high agreement with human judgments (r = 0.6; Pincombe, 2004), I focused on the most extreme errors of LSA (i.e., extreme misses and false positives) determined based on the *degree* of error rather than focusing on how successfully it can make binary judgments of document relatedness (i.e., precision and recall).

Misses prevent human readers from reviewing relevant documents. False positives would waste human readers' time and effort. While false positives at least give a chance for human readers to decide whether the retrieved documents are relevant or not, misses completely eliminate the chance to review potentially relevant documents. Therefore, one may consider

misses worse than false positives. However, another way of viewing the issue is that misses and false positives are actually two sides of the same coin. An algorithm retrieves relevant documents to a given query, and then presents them in an order of their relevance. If it misses some relevant documents (i.e., presents highly relevant documents after it presents less relevant documents), it is because it false positively judged other (less relevant) documents to be more related to a query than the missed ones, and vice versa. Hence, in the present study, I focused on both kinds of errors, with the goal of improving the overall document relatedness judgment by LSA.

Misses by LSA in Document Relatedness Judgments

In the current study, missed document pairs were operationally defined as having an absolute *z*-score difference of 1 or greater between humans and LSA, with relatively high human ratings and relatively low LSA ratings. A potential reason for misses by LSA is that LSA misses word relatedness in the document pairs (see Chapters 4). However, LSA may miss something else (rather than word relatedness) that led humans to judge the document pairs to be related. For example, humans can judge two documents to be related based on the degree of overlap of their writing structures (e.g., both documents describe a certain event and then describe reactions of people to the event), or their hierarchical relationship (e.g., one document is explaining the meaning of a concept and the other document is showing an example of the concept). The vector-space model, or its variant LSA, are not designed to capture such higher-level relationship between documents. They were designed to capture the semantic relatedness between documents, which is operationally defined as the *degree of overlap of words* in documents.

Therefore, researchers who attempt to develop an AI algorithm that judges document relatedness as humans do first need to determine what are the users' definitions of *relatedness*.

For example, users may want to find document pairs with a similar writing structure rather than a similar topic (e.g., researchers who want to find documents that have the typical journal-article structure from internet). Further work is needed to examine various criteria of document relatedness or similarity judgment used by human readers.

False Positives by LSA in Document Relatedness Judgments

Since we already know much about how LSA works (i.e., LSA judges document relatedness based on the word relatedness between documents), the reason LSA assigns relatively high ratings to false positive document pairs is clear (i.e., false positive documents have a high degree of overall word relatedness). Regarding the false positives, however, the problem is that we do not fully understand why humans judge them to be unrelated even though they have relatively high degree of word relatedness. Although I proposed one potential explanation (i.e., the key-features overlap model), and showed that it has some validity, there may be other reasons. For example, even if a document pair has a high degree of keyword relatedness as well as a high degree of overall word relatedness, the document pair can be judged to be unrelated by humans, if the two documents have completely different intentions of writing. For example, both Aesop's fable of *the fox and sour grapes* and an imaginary scientific article regarding foxes' diet may have the same keywords, such as fox, eat, and grape along with numerous overlapping non-keywords. However, humans may judge them to be unrelated due to the completely different intentions of the authors. Further studies are needed examining how humans judge the relatedness of false positive document pairs.

Implications of the Current Study

There are two main implications of the current study. One is about human similarity judgment in general and the other is about improving the document relatedness judgments of AI algorithms.

Human Similarity Judgments Based on Key Features

One of the main findings of the current study is that humans judge document relatedness by placing more weight on the relatedness of keywords, and less weight on the relatedness of non-keywords. The key-features overlap model was supported by the demonstration that the relatedness of keywords could capture the differences of human document relatedness ratings, whereas the relatedness of all words could not (Chapter 5). Notably, the key-features overlap model may be applicable to human similarity judgments in general. That is, when humans make similarity judgments, they may put more weight on the degree of overlap among key features. Depending on the object, key features may differ. For example, when the objects are analogical sentences, the key feature could be the structure of the two analogies rather than the individual words within them, as Gentner (1983) suggested. When the objects are faces, the similarity of the eyes, for example, may play a greater role in judging the similarity of the faces than the similarity of noses.

Further studies are needed examining whether the key-features overlap model is applicable to other stimuli domains. For example, some on-line dating sites suggest potentially well-matching mates based on the similarity of users' text profiles. However, users may (implicitly) put more weight on certain features of a profile with which they hope their potential mate has more similarity to their own. Then, as the highly-weighted features (key features) are more similar to each other between users, their satisfaction with the suggested mates may be greater than when only less important features are matching well. Similar issues exist in

consumers' behavior of choosing products or brands (Park, Milberg, & Lawson, 1991). Although the key-feature overlap model seems to be intuitively valid, the actual functions between the levels of weight of individual features, the similarity level of those features, and how the function affects the final similarity judgments of the objects deserve further studies.

The Effect of Good Background Documents on LSA Performance

The current study showed the beneficial effects of having good background documents on LSA's linguistic performance. The goodness of background documents is determined by both the number and quality of the documents. More background documents are beneficial under the condition that they contain relevant information to the target documents. In particular, according to the key-features overlap model, the background documents should contain sufficient information regarding the keywords of the target documents. Information regarding proper nouns (e.g., names of people or regions) among keywords may not be sufficiently represented in the background documents. Therefore, having dictionary information regarding them could improve LSA's performance. Further studies are needed to demonstrate whether such suggestions derived from the understanding of human cognition can in fact enhance the performance of linguistic AI algorithms.

Table 1.

Document ID numbers, types of misjudgment, and averaged human ratings of the 22 candidates of misjudged document pairs. Final misjudged document pairs selected for further study were shaded in gray.

First Document	Second Document	Mismatch Type	Original Relatedness Scores (Pincombe, 2004)	Current Relatedness Scores
12	24	Miss	3.5	4.3
18	37	Miss	2.9	3.5
22	49	Miss	2.7	3.3
40	43	Miss	4.3	3.1
19	20	Miss	3.2	2.8
3	33	Miss	2.7	2.8
4	29	Miss	3.1	2.5
21	27	Miss	3.0	2.3
34	41	Miss	3.5	2.2
1	22	Miss	2.6	1.8
10	17	False Positive	1.6	2.9
15	19	False Positive	1.2	2.4
23	35	False Positive	1.2	2.3
16	22	False Positive	1.2	2.1
42	45	False Positive	1.3	2.0
28	30	False Positive	1.2	2.0
2	43	False Positive	1.1	1.6
7	49	False Positive	1.1	1.5
24	39	False Positive	1.1	1.4
27	41	False Positive	1.0	1.3
18	31	False Positive	1.1	1.2
6	37	False Positive	1.0	1.1

Table 2.

Example of the missed and false positive document pairs.

Mismatch Type	First Document	Second Document
Miss	40. The real level of world inequality and environmental degradation may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by the Guardian. It includes new estimates that the world lost almost 10% of its forests in the past 10 years; that carbon dioxide emissions leading to global warming are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020.	43. Pope John Paul II urged delegates at a major U.N. summit on sustainable growth on Sunday to pursue development that protects the environment and social justice. In comments to tourists and the faithful at his summer residence southeast of Rome, the pope said God had put humans on Earth to be his administrators of the land, "to cultivate it and take care of it." "In a world ever more interdependent, peace, justice and the safekeeping of creation cannot but be the fruit of a joint commitment of all in pursuing the common good," John Paul said.
	Human ratings of 4.3 (*z*-score of 3.84, Pincombe, 2004), 3.1 (current study) LSA rating of 0.11 (*z*-score of 0.23)	
False positive	7. Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden's terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa'ida, including bin Laden. The money enabled al-Qa'ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia.	49. Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year.
	Human ratings of 1.1 (*z*-score of - 0.76, Pincombe, 2004), 1.5 (current study) LSA rating of 0.18 (*z*-score of 0.8)	

Table 3.

Average relatedness ratings of all word pairs in misjudged document pair.

First Document	Second Document	Mismatch Type	Average Word Relatedness	Average Keyword Relatedness	Original Document Relatedness Scores (Pincombe, 2004)	Current Relatedness Scores
12	24	Miss	3.16	3.74	3.5	4.3
18	37	Miss	2.93	3.17	2.9	3.5
22	49	Miss	3.16	3.15	2.7	3.3
40	43	Miss	2.99	3.07	4.3	3.1
19	20	Miss	2.87	2.81	3.2	2.8
3	33	Miss	2.91	2.82	2.7	2.8
2	43	False Positive	2.92	2.59	1.1	1.6
7	49	False Positive	3.1	2.85	1.1	1.5
24	39	False Positive	2.71	2.65	1.1	1.4
27	41	False Positive	2.71	2.42	1.0	1.3
18	31	False Positive	2.75	2.81	1.1	1.2
6	37	False Positive	2.92	2.54	1.0	1.1

Table 4.

The 30 most highly related word pairs by humans and LSA from a false positive document pair (document pair of 7 and 49).

Human		LSA		
Related word pair	Human ratings	Related word pair	LSA ratings	*z*-scores
dollar~money	5.00	design~document	0.71	6.87
job~money	5.00	increase~rise	0.59	5.61
angrily~attack	4.90	paid~worker	0.44	3.97
plan~target	4.83	effect~target	0.39	3.38
increase~profit	4.80	group~work	0.39	3.37
money~profit	4.80	effect~increase	0.36	3.08
cost~lawsuit	4.78	disclosure~profit	0.35	2.89
job~meet	4.75	disclosure~financial	0.34	2.83
agreement~plan	4.73	commonwealth~deal	0.30	2.41
job~paid	4.73	australia~target	0.29	2.32
meet~work	4.70	cost~paid	0.29	2.30
meet~worker	4.70	cut~fund	0.29	2.23
angrily~lawsuit	4.67	group~plan	0.28	2.14
financial~money	4.62	australian~increase	0.28	2.12
cost~paid	4.60	allege~angrily	0.27	2.10
increase~largest	4.60	cost~fund	0.27	2.06
paper~work	4.60	leader~plan	0.27	2.05
paper~worker	4.60	cut~target	0.27	2.03
cut~paper	4.58	bank~increase	0.27	2.03
cost~money	4.55	australia~meet	0.27	2.01
effect~lawsuit	4.55	australia~increase	0.27	2.01
bank~money	4.50	australia~enable	0.26	1.96
document~work	4.50	camp~control	0.26	1.95
document~worker	4.50	control~group	0.26	1.94
dollar~paper	4.50	report~senior	0.25	1.88
financial~paid	4.50	control~tension	0.25	1.87
fiscal~lawsuit	4.50	deal~effect	0.25	1.87
leader~worker	4.50	disclosure~fiscal	0.25	1.86
angrily~terror	4.45	senior~work	0.25	1.82
plan~secret	4.44	increase~profit	0.25	1.81

Figure 1.

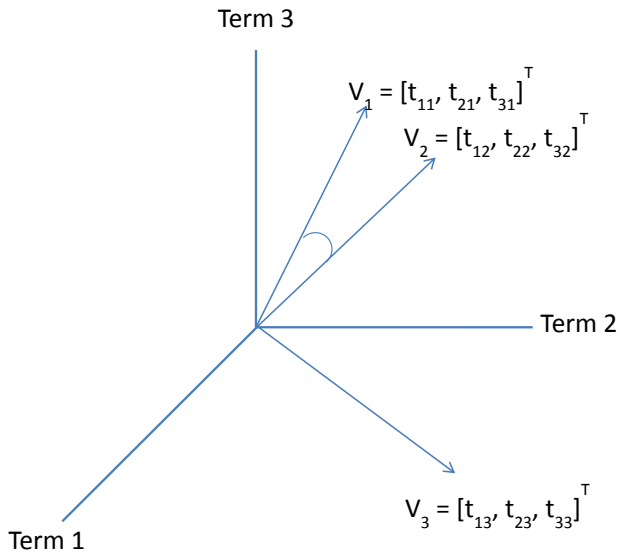

Figure 1. Example of document vectors depicted on a three dimensional word-space. The first two document vectors (V_1 and V_2) are close to each other, while the third one (V_3) is far from them. Such geographic relation suggests that the first two documents have similar words and accordingly a similar topic, while the third one (V_3) has a different topic.

Figure 2.

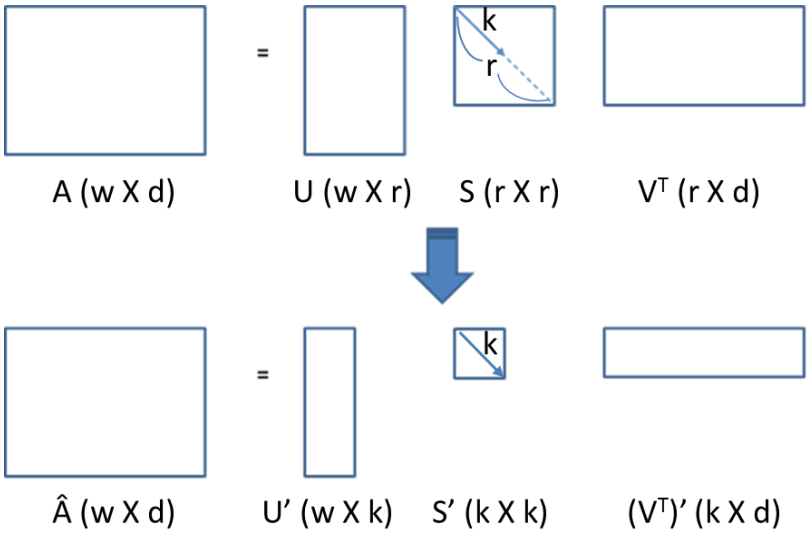

Figure 2. Full SVD (upper panel) and reduced SVD (lower panel).

Figure 3.

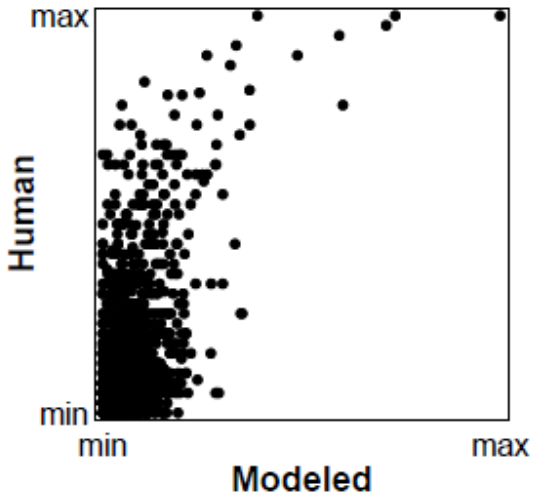

Figure 3. The relationship between humans and the vector-space model (with common-features measure) with respect to their document relatedness judgments. Each dot represents a document pair. The X-axis shows the range of similarity scores made by the model and the Y-axis shows the range of human ratings.

Figure 4.

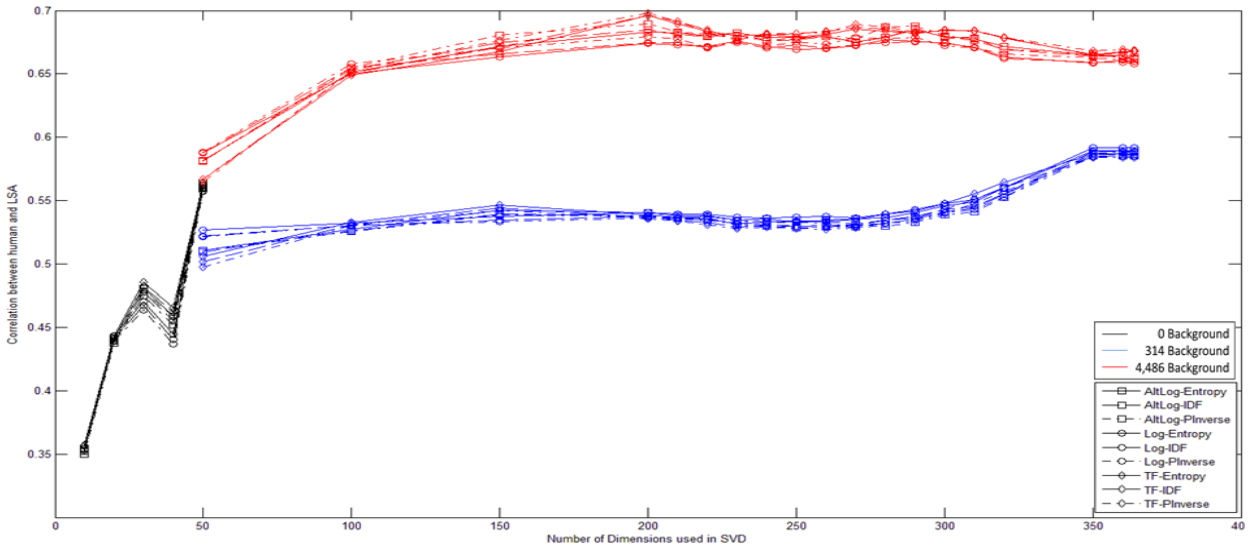

Figure 4: LSA-parameter manipulation results. The black, blue, and red lines indicate the three different background document conditions of 0, 314, and 4,486, respectively. The markers (square, circle, and diamond) indicate the three different local weighting schemes, and the three different line styles (solid, dashed, and dash-dot) indicate the three global weighting schemes (see legend). The X-axis shows the number of dimensions used in SVD, and the Y-axis shows the correlation (*r*) between human document relatedness ratings and LSA-cosine scores across the 1,225 news article pairs.

Figure 5.

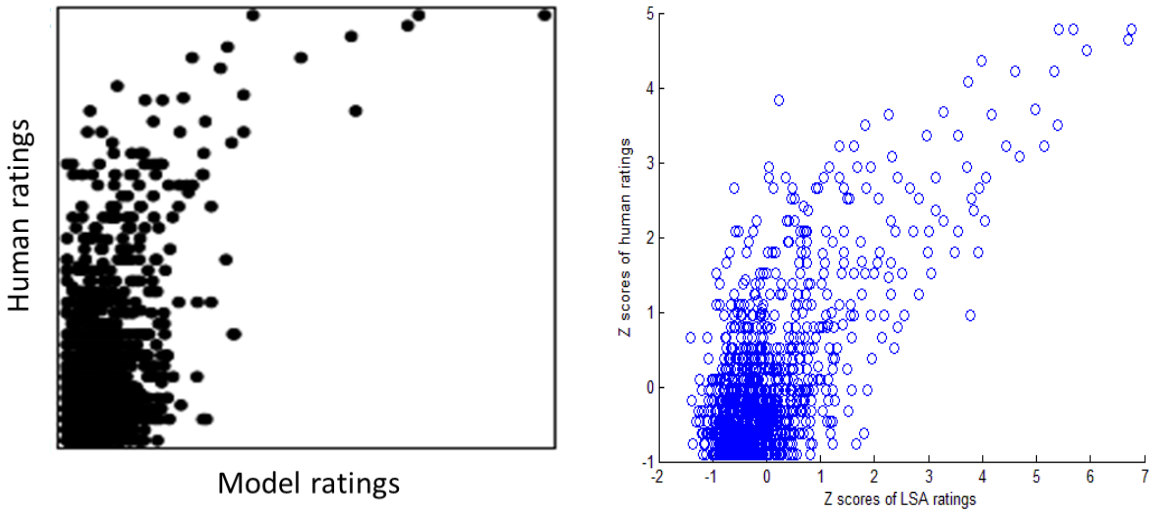

Figure 5. Left panel: relationship between the ratings of humans and a computational model (a vector space model with common-features measure), across the 1,225 news article pairs as reported in Lee et al. (2005). Right panel: relationship between the normalized ratings of humans and optimal-parameter-based normalized LSA-cosine scores, for the same 1,225 document pairs. Each circle represents a document pair, the X-axis shows the model scores, and the Y-axis shows the human ratings.

Figure 6.

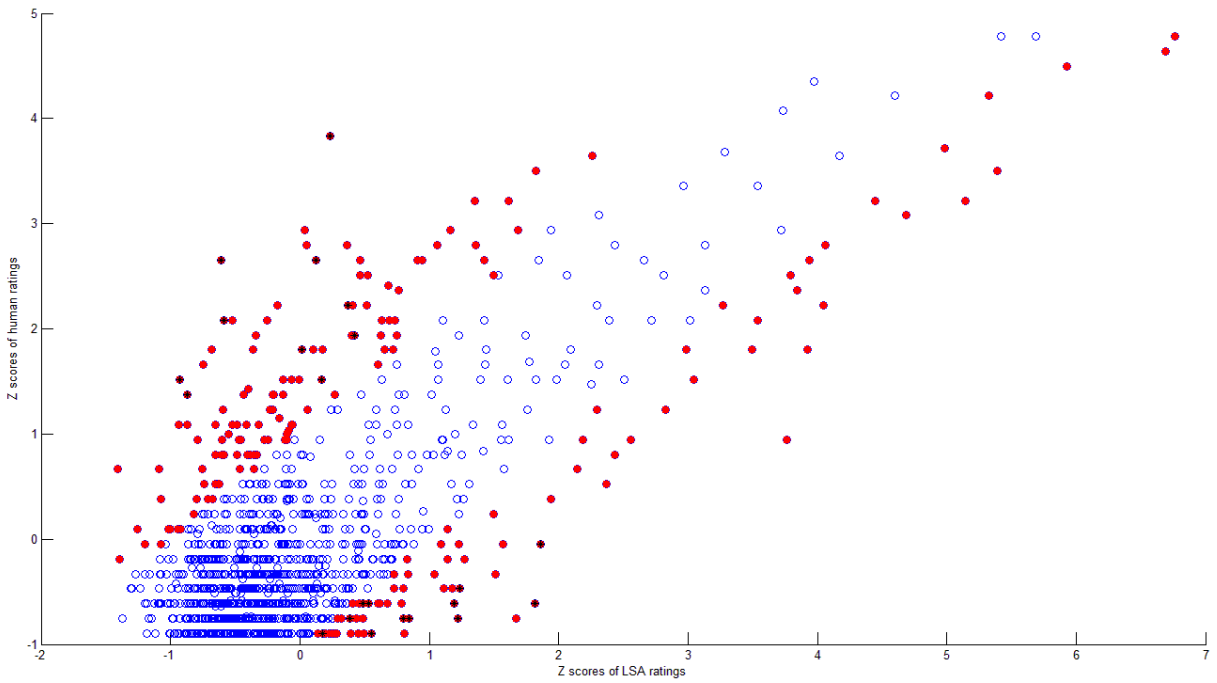

Figure 6. Two-hundred and three candidates of misjudged document pairs. The scatter plot is the same as the right panel of Figure 5. Document pairs with absolute *z*-score differences of 1 or greater were selected as candidates of the misjudged document pairs and marked in red. Among them, 10 candidates of missed document pairs and 12 candidates of false positive document pairs were selected for further study (marked with black dots inside of circles).

Figure 7.

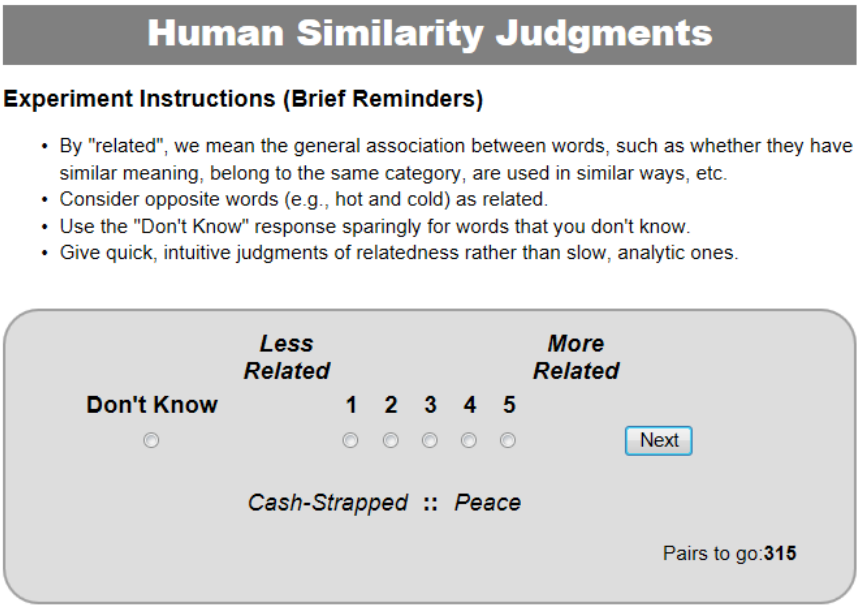

Figure 7. Sample trial from the word relatedness rating experiment (Chapter 5).

Figure 8.

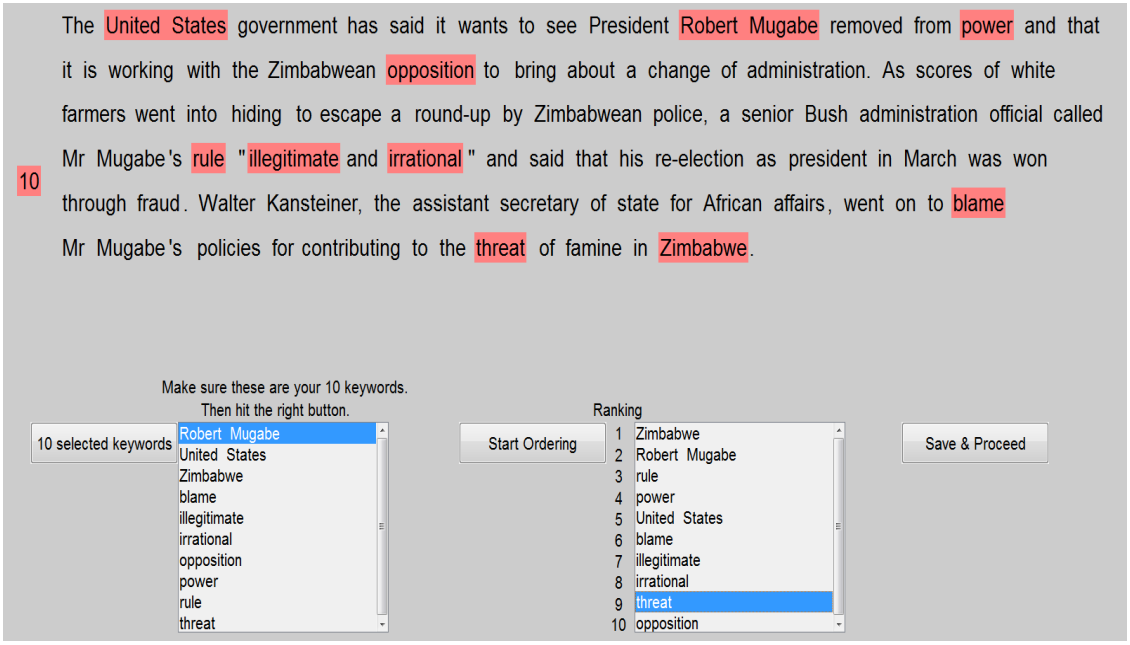

Figure 8. Sample trial from the key word finding experiment (Chapter 6).

Figure 9.

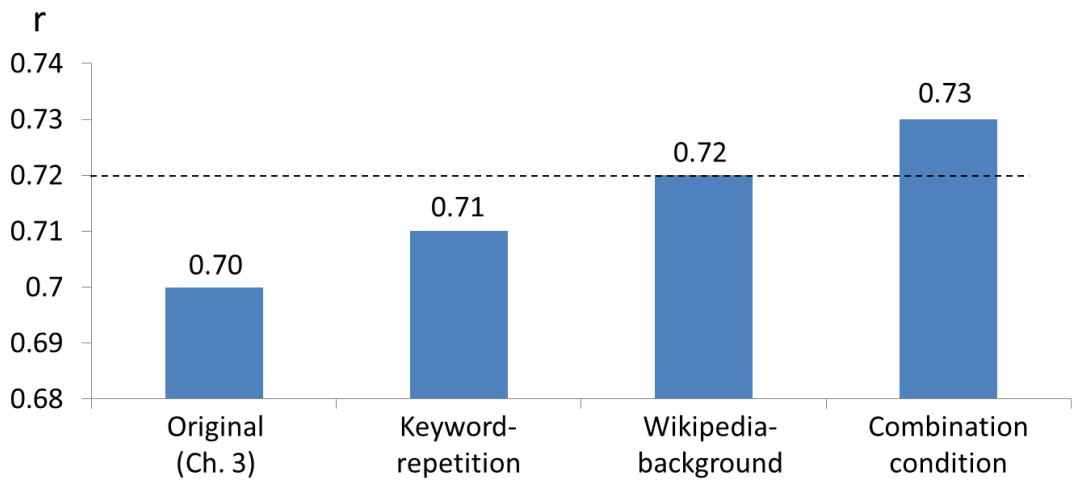

Figure 9. Agreement (*r*) between humans and the three modified versions of LSA (the keyword-repetition, the Wikipedia-background, and the combination condition) in Chapter 7 (as well as the result from Chapter 3). The dashed line shows the correlation of 0.72 obtained from Gabrilovich and Markovitch (2007).

Appendix A. Various pre-processing procedures

Pre-processing	Functions
Stopword removal	Removing stopwords (e.g., the).
Filtering	Removing special characters (e.g., %, @).
Pruning	Removing words that occur too rarely or too frequently. A priori threshold is applied (e.g., a word should occur at least twice across all documents to be kept in the word-by-document matrix).
Stemming	Removing of any attached suffixes and prefixes from the word to yield the word stem. Porter's algorithm is widely used (Porter, 1980).
Tokenization	Although the typical unit of a dimension in the vector-space model is a word, it is not necessarily a word. It can be phrases or any other combinations of words employed to identify the contents of a document.

Note. The above pre-processing terms were adopted from Andrews and Fox (2007).

Appendix B. Fifty Australian news articles used in Pincombe (2004) as well as in the current study

1. The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist. (80 words)
2. Cash-strapped financial services group AMP has shelved a $400 million plan to buy shares back from investors and will raise $750 million in fresh capital after profits crashed in the six months to June 30. Chief executive Paul Batchelor said the result was "solid" in what he described as the worst conditions for stock markets in 20 years. AMP's half-year profit sank 25 per cent to $303 million, or 27c a share, as Australia's largest investor and fund manager failed to hit projected 5 per cent earnings growth targets and was battered by falling returns on share markets. (98 words)
3. The United States government has said it wants to see President Robert Mugabe removed from power and that it is working with the Zimbabwean opposition to bring about a change of administration. As scores of white farmers went into hiding to escape a round-up by Zimbabwean police, a senior Bush administration official called Mr Mugabe's rule "illegitimate and irrational" and said that his re-election as president in March was won through fraud. Walter Kansteiner, the assistant secretary of state for African affairs, went on to blame Mr Mugabe's policies for contributing to the threat of famine in Zimbabwe. (98 words)
4. A radical armed Islamist group with ties to Tehran and Baghdad has helped al-Qaida establish an international terrorist training camp in northern Iraq, Kurdish officials say. Intelligence officers in the autonomous Kurdish region of Iraq told the Guardian that the Ansar al-Islam (supporters of Islam) group is harbouring up to 150 al-Qaida members in a string of villages it controls along the Iraq-Iran border. Most of them fled Afghanistan after the US-led offensive, but officials from the Patriotic Union of Kurdistan (PUK), which controls part of north-east Iraq, claim an "abnormal" number of recruits are making their way to the area from Jordan, Syria and Egypt. (106 words)

5. Washington has sharply rebuked Russia over bombings of Georgian villages, warning the raids violated Georgian sovereignty and could worsen tensions between Moscow and Tbilisi. "The United States regrets the loss of life and deplores the violation of Georgia's sovereignty," White House spokesman Ari Fleischer said. Mr Fleischer said US Secretary of State Colin Powell had delivered the same message to his Russian counterpart but that the stern language did not reflect a sign of souring relations between Moscow and Washington. (80 words)
6. A gay former student of a Melbourne Christian school is taking legal action under equal opportunity legislation, claiming the school discriminated against him because of his sexuality. Tim, 16, alleged a staff member at Hillcrest Christian College in Berwick told him he "had the devil in him", and constant bullying by students prompted the principal to tell him to hide his sexuality. He left the school several weeks ago and is continuing Year 10 by distance education after he said homophobic bullies threw rocks at his head, spat on him, called him names and slashed his belongings. (97 words)
7. Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden's terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa'ida, including bin Laden. The money enabled al-Qa'ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia. (97 words)
8. Palestinian hired gun Abu Nidal, whose violent death was reported last week from Baghdad, was murdered on the orders of Iraqi President Saddam Hussein after refusing to train al-Qa'ida fighters based in Iraq, reports said yesterday. Iraqi intelligence chief Taher Jalil Habbush said last Wednesday Abu Nidal had shot and killed himself after being discovered living illegally in Baghdad and facing interrogation for anti-Iraqi activities. But Western diplomats believe the radical militant was killed for refusing to reactivate his international terrorist network. (82 words)
9. Hunan province remained on high alert last night as thunderstorms threatened to exacerbate the flood crisis, now entering its fifth day and with 108 already dead and hundreds of thousands evacuated. On the flood frontline at Dongting Lake, the water level peaked at just

under 35m on Saturday night, then eased about 3cm during the day under a hot sun, with temperatures reaching 35C. But with the lake still brimming at dangerously high levels, and spilling over the top of its banks in some places, locals were fearful that a thunderstorm and high winds forecast to hit the region last night would damage the dikes. About 1800km of dikes around the lake are all that stand between 10 million people in the surrounding farmland and disaster. (126 words)

10. A U.S.-British air raid in southern Iraq left eight civilians dead and nine wounded, the Iraqi military said Sunday. The military told the official Iraqi News Agency that the warplanes bombed areas in Basra province, 330 miles south of Baghdad. The U.S. Central Command in Florida said coalition aircraft used precision-guided weapons to strike two air defense radar systems near Basra "in response to recent Iraqi hostile acts against coalition aircraft monitoring the Southern No-Fly Zone." (76 words)
11. Iraq and Russia are close to signing a $40 billion economic cooperation plan, Iraq's ambassador said Saturday, a deal that could put Moscow at odds with the United States as it considers a military attack against Baghdad. The statement by Ambassador Abbas Khalaf came amid indications that Russia, despite its strong support for the post-Sept. 11 antiterrorism coalition, is maintaining or improving ties with Iran and North Korea, which together with Iraq are the countries President Bush has labeled the "axis of evil." (83 words)
12. U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials. (67 words)
13. Drug squad detectives have asked the Police Ombudsman to investigate the taskforce that is examining allegations of widespread corruption within the squad. This coincides with the creation of a special unit within the taskforce to track the spending of at least 10 serving and former squad members. The corruption taskforce, codenamed Ceja, will check tax records and financial statements in a bid to establish if any of the suspects have accrued unexplained wealth over the past seven years. But drug squad detectives have countered with their own

set of allegations, complaining to the ombudsman that the internal investigation is flawed, biased and over-zealous. (103 words)

14. Queensland senator Andrew Bartlett has launched a last-minutebid to rescue the Australian Democrats from a split that threatens to destroy the party. With nominations for the party leadership to close on Wednesday night, Senator Bartlett met last night with deputy leader Aden Ridgeway to offer him a place on a unity ticket and set up a reform process to begin healing the party's wounds. Party sources said Senator Ridgeway, who turned against former leader Natasha Stott Despoja, is still expected to contest the leadership against one of her two supporters: Senator Bartlett or Brian Greig, installed as interim leader by the party's executive last Thursday. (105 words)
15. Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline. (90 words)
16. The Bush administration has drawn up plans to escalate the war of words against Iraq, with new campaigns to step up pressure on Baghdad and rally world opinion behind the US drive to oust President Saddam Hussein. This week, the State Department will begin mobilising Iraqis from across North America, Europe and the Arab world, training them to appear on talk shows, write opinion articles and give speeches on reasons to end President Saddam's rule. (75 words)
17. Beijing has abruptly withdrawn a new car registration system after drivers demonstrated "an unhealthy fixation" with symbols of Western military and industrial strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man's choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels. (73 words)
18. The United Nations World Food Program estimates that up to 14 million people in seven countries - Malawi, Mozambique, Zambia, Angola, Swaziland, Lesotho and Zimbabwe - face

death by starvation unless there is a massive international response. In Malawi, as many as 10,000 people may have already died. The signs of malnutrition - swollen stomachs, stick-thin arms, light-coloured hair - are everywhere. (62 words)

19. In Malawi, as in other countries in the region, AIDS is making the effects of the famine much worse. The overall HIV infection rate in Malawi is 19 per cent, but in some areas up to 35 percent of people are infected. A significant proportion of the young adult population is too sick to do any productive work. Malnutrition causes people to succumb to the disease much more quickly than they do in the West, and hunger forces women into prostitution in order to feed their families, making them more vulnerable to contracting the disease. Life expectancy has been reduced to 40 years. (103 words)
20. The United Nations was determined that its showpiece environment summit - the biggest conference the world has ever witnessed - should be staged in Africa. The venue, however, could not be further removed from the grim realities of life in the rest of Africa. Johannesburg's exclusive and formerly whites-only suburb of Sandton is the wealthiest neighbourhood in the continent. Just a few kilometres from Sandton begins the sprawling Alexandra township, where nearly a million people live in squalor. Organisers of the conference, which begins today, seem determined that the two worlds should be kept as far apart as possible. Tight security surrounds Sandton's convention centre and five-star hotels, where world leaders will debate poverty, the environment and sustainable development while enjoying lavish hospitality. (122 words)
21. The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle. (94 words)
22. The Federal Government says changes announced today to the work for the dole scheme will benefit participants and taxpayers. Federal Employment Services Minister Mal Brough says that from July 1 those taking part in work for the dole will be able to perform extra hours to complete their mutual obligation more quickly to access training credits. (61 words)

23. The biowarfare expert under scrutiny in the anthrax attacks declared, "I am not the anthrax killer," and lashed out today against Attorney General John Ashcroft for calling him a "person of interest" in the investigation. For the second time in two weeks, the scientist went before a throng of reporters outside his lawyer's office to profess his innocence and decry the attention from law enforcers that he contends has destroyed his life. (72 words)
24. China said Sunday it issued new regulations controlling the export of missile technology, taking steps to ease U.S. concerns about transferring sensitive equipment to Middle East countries, particularly Iran. However, the new rules apparently do not ban outright the transfer of specific items - something Washington long has urged Beijing to do. (54 words)
25. Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry. (57 words)
26. An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing. (61 words)
27. How did 2,300 allegedly unregistered missile warheads come to be stored on a Canadian businessman's anti-terrorism training facility in New Mexico? U.S. and Canadian officials are still trying to figure that out, but one security expert says the mystery is a "chilling" one. David Hudak, 41, was arrested in the United States more than a week ago when, according to court documents, agents searching his property found the warheads stored in crates that were marked "Charge Demolition." (77 words)
28. The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed "and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court," the official Saudi Press Agency quoted an unidentified ministry official as saying. (65 words)
29. Sri Lanka's government will lift a four-year ban on Tamil Tiger rebels on Sept. 6, paving the way for peace talks with the insurgents scheduled for later that month in Thailand, a

government minister said Saturday. "We will lift the ban as promised," Minister for Rehabilitation Jayalath Jayawardena told The Associated Press. The lifting of the ban is one of the key rebel conditions for resuming peace negotiations with the government after a hiatus of more than seven years. (79 words)

30. A man accused of making hidden-camera footage up the skirts of women also made child pornography of the worst kind, featuring the rape of children as young as 6, police said Friday. The latest allegations suggest there's nothing humorous about voyeurs who some may perceive to be making secret videos as a joke, Staff-Insp. Gary Ellis said. "Approximately 20 per cent of voyeurs have also committed sexual assault or rape," Ellis said, reading from a recently released federal government report on criminal voyeurism. (83 words)
31. Police are combing through videotapes trying to spot the gunman dressed in black who shot a 30-year-old man to death at a downtown massage parlour. The victim was hit in the stomach and upper body and died about 3 1/2 hours later in hospital. The woman was not hurt. Police urged business owners to turn over any security-camera videotapes they might have that recorded people on the street at the time. Several such videos are now being reviewed. (78 words)
32. The Federal Government did not regret its actions 12 months on from the Tampa asylum seeker crisis, Small Business Minister Joe Hockey said today. Mr Hockey said the Government was not embarrassed by the Tampa issue, which began on August 27 of last year when the captain of the Norwegian cargo ship rescued more than 400 asylum seekers from an Indonesian ferry north of Christmas Island. (66 words)
33. At least three Democrats are considering splitting from the party while no-one has yet nominated to contest the leadership. Three of the "gang of four" senators who ousted Natasha Stott Despoja from the leadership are considering forming a new "progressive centre" party in the fallout from last week's turmoil. This would leave the Democrats with a rump of three or four members. West Australian Senator Andrew Murray said yesterday unless the Democrats left wing gave ground the party would split. (80 words)
34. A young humpback whale remained tangled in a shark net off the Gold Coast yesterday, despite valiant efforts by marine rescuers. With its head snared by the net and an anchor rope wrapped around its tail, the stricken whale was still swimming but hopes for its survival were

fading. A second rescue attempt was planned for dawn today after rescuers braved heavy seas, strong wind and driving rain to try to free the whale. (74 words)

35. Prince William has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. William and Prince Harry are so devastated by the treachery of Ken Wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives. (63 words)
36. The spectre of Osama bin Laden rose again today, urging Afghans to launch a new Jihad, or holy war, and predicting the fall of the United States, in a hand-written "letter" posted on an Islamic website. There was no hard proof that the scruffy missive was genuine, but IslamOnline.net said it had been received by their correspondent in Jalalabad, eastern Afghanistan, from an Afghan source who asked to remain anonymous. The source claimed it was the "most recent letter" from the world's most wanted man. (85 words)
37. The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African President Thabo Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US President George W Bush was threatening to overshadow the summit. (65 words)
38. Robert Mugabe strengthened his hold on the Zimbabwean government yesterday by retaining the most combative hardliner ministers in a cabinet shuffle which offered little hope of a moderation of the land seizures and other policies that have kept Zimbabwe in crisis and brought international condemnation. (51 words)
39. They dress in black and disguise their identities with bandannas and sunglasses. Their logo is an image of the Southern Cross constellation, superimposed with a pair of crossed boomerangs, which resembles a swastika. The Blackshirts are former husbands aggrieved by their treatment at the hands of their ex-wives and the courts, who regard themselves as the vanguard of a "men's rights" movement in Australia and say that their actions will be remembered as marking a turning-point in history. (78 words)
40. The real level of world inequality and environmental degradation may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by the Guardian. It includes new estimates that the world lost almost 10% of its

forests in the past 10 years; that carbon dioxide emissions leading to global warming are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020. (93 words)

41. Researchers conducting the most elaborate wild goose chase in history are digesting the news that a bird they have tracked for over 4,500 miles is about to be cooked. Kerry, an Irish light-bellied Brent goose, was one of six birds tagged in Northern Ireland in May by researchers monitoring the species' remarkable migration. Last week, however, he was found dead in an Inuit hunter's freezer in Canada, still wearing his £3,000 satelite tracking device. Kerry was discovered by researchers on the remote Cornwallis Island. They picked up the signal and decided to try to find him. (96 words)
42. Russia defended itself against U.S. criticism of its economic ties with countries like Iraq, saying attempts to mix business and ideology were misguided. "Mixing ideology with economic ties, which was characteristic of the Cold War that Russia and the United States worked to end, is a thing of the past," Russian Foreign Ministry spokesman Boris Malakhov said Saturday, reacting to U.S. Defense Secretary Donald Rumsfeld's statement that Moscow's economic relationships with such countries sends a negative signal. (77 words)
43. Pope John Paul II urged delegates at a major U.N. summit on sustainable growth on Sunday to pursue development that protects the environment and social justice. In comments to tourists and the faithful at his summer residence southeast of Rome, the pope said God had put humans on Earth to be his administrators of the land, "to cultivate it and take care of it." "In a world ever more interdependent, peace, justice and the safekeeping of creation cannot but be the fruit of a joint commitment of all in pursuing the common good," John Paul said. (96 words)
44. The Russian defense minister said residents shouldn't feel threatened by the growing number of Chinese workers seeking employment in the country's sparsely populated Far Eastern and Siberian regions. There are no exact figures for the number of Chinese working in Russia, but estimates range from 200,000 to as many as 5 million. Most are in the Russian Far East, where they arrive with legitimate work visas to do seasonal work on Russia's low-tech, labor-intensive farms. (75 words)
45. Australian spies listened to conversations between Norway's ambassador and its foreign office during the Tampa crisis, a soon to be published book will reveal. Phone calls were

tapped by the Defence Signals Directorate when Norwegian ambassador Ove Thorsheim visited the freighter during the stand-off. A book, Tampa, to be published in Norway in October, recounts the events which triggered Australia's Pacific Solution and transformed Tampa Captain Arne Rinnan into a homeland hero. (72 words)

46. Batasuna, a political party that campaigns for an independent Basque state, faces a double blow today: the Spanish parliament is expected to vote overwhelmingly in favour of banning the radical group, while a senior investigative judge is poised to suspend Batasuna's activities on the grounds that they benefit Eta, the outlawed Basque separatist group. (56 words)
47. The river Elbe surged to an all-time record high Friday, flooding more districts of the historic city of Dresden as authorities scrambled to evacuate tens of thousands of residents in the worst flooding to hit central Europe in memory. In the Czech Republic, authorities were counting the cost of the massive flooding as people returned to the homes and the Vlava river receded, revealing the full extent of the damage to lives and landmarks. (74 words)
48. The European Parliament is spoiling for a fight with Israel. It has voted to review the EU's diplomatic links with the Jewish state, to impose an arms embargo and to threaten wider trade sanctions. Many MEPs want to go further and dispatch a European military force to the region in order to "protect the Palestinian people". (58 words)
49. Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year. (79 words)
50. Labor needed to distinguish itself from the Government on the issue of asylum seekers, Greens leader Bob Brown has said. His Senate colleague Kerry Nettle intends to move a motion today - on the first anniversary of the Tampa crisis - condemning the Government over its refugee policy and calling for an end to mandatory detention. "We Greens want to bring the Government to book over its serial breach of international obligations as far as asylum seekers in this country are concerned," Senator Brown said today. (86 words)

References

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behavior* (2nd ed.). New York, NY: John Wiley and Sons.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review, 37*, 573–595.

Bullinaria, J., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*, 510–526.

Connor, J. (2012, August 8). Scholar updates: Making new connections [Web log post]. Retrieved from http://googlescholar.blogspot.com/2012/08/scholar-updates-making-new-connections.html.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers, 23*, 229–236.

Dumais, S. T. (2003). Data-driven approaches to information access. *Cognitive Science, 27*, 491–524.

Foltz, P. W., Britt, M. A., & Perfetti, C. A (1995). Measuring text influence on learning history. *Paper presented at the 5th Annual Winter Text Conference*, Jackson, WY.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *Proceedings of the 11th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval. 465–480. Grenoble, France: ACM. doi:10.1145/62437.62487.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 1606–1611.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Kang, S. (2003). Keyword-based Document Clustering. *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, Sappro, Japan, 11*, 132–137.

Kiley, R. (2010, March 1). European research funders throw weight behind UK open access repository [Web log post]. Retrieved from http://blog.europepmc.org/2010/03/european-research-funders-throw-weight.html

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Lee, M.D., Pincombe, B.M., Welsh, M.B., 2005. An empirical evaluation of models of text document similarity. In: Bara, B.G., Barsalou, L., Bucciarelli, M. (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, Stresa, Italy. Cognitive Science Society, Austin, TX, pp. 1254–1259.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788–791. doi:10.1038/44565.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems, 13,* 556–562.

Lintean, M., Moldovan, C., Rus, V., & McNamara D. S. (2010). The role of local and global weighting in assessing the semantic similarity of texts using Latent Semantic Analysis. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*. Daytona Beach, FL.

Luger, G, F. (2009). *Artificial intelligence: Structures and strategies for complex problem solving* (6th ed.). Boston, MA: Pearson Education.

Madsen, R. E., Hansen, L. K., & Winther, O. (2003). Singular value decomposition and principal component analysis. *ISP technical report*.

Matsuo, Y., & Ishizuka, M. (2003) Keyword extraction from a single document using word co-occurrence statistical information. *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference*, 392-396.

McCorduck, P. (2004). *Machines who think* (2nd ed.). Natick, MA: A. K. Peters.

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. *In Proceedings of the EuroConference Recent Advances in Natural Language Processing*, 187–193.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56,* 132–143.

Park, W. C, Milberg, S., & Lawson, R. (1991). Evaluation of brand extensions: the role of product feature similarity and brand concept consistency. *Journal of Consumer Research, 18*, 185–193.

Pincombe, B. M. (2004). Comparison of human and LSA judgments of pairwise document similarities for a news corpus. Technical report. No. DSTO-RR-0278. Adelaide, Australia: Australian Defense Science and Technology Organization (DSTO), Intelligence, Surveillance and Reconnaissance Division. Available at: http://hdl.handle.net/1947/3334.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*, 37-63.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043–1056.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). New Jersey, NY: Prentice Hall.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill computer science series. New York, NY: McGraw-Hill.

Polettini, N. (2004). *The vector space model in information retrieval-term weighting problem.* Department of information and communication Technology, University of Trento, Italy.

Sahlgren, M. (2005). An introduction to random indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*, 1–47.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: nineteen results all computer vision researchers should know about. *Proceedings of the IEEE, 94,* 1948–1962.

Stone, B., Dennis, S., & Kwantes, P. (2011). Comparing methods for single paragraph similarity analysis, *Topics in Cognitive Science, 3*, 92–122.

Toffler, A. (1970). *Future shock*. New York, NY: Random House.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Zeimpekis, D. & Gallopoulos, E. (2012). Text to matrix generator (Version 4) [Computer Toolbox]. Available from http://scgroup6.ceid.upatras.gr:8000/wiki/index.php/Main_Page.