7-18-2013

Evaluation: Measurements of Differences between Semantic Spaces

Xudong Huang

EVALUATION: MEASUREMENTS OF DIFFERENCES BETWEEN SEMANTIC SPACES

by

Xudong Huang

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

August 2013

ACKNOWLEDGMENTS

I would like to first thank my major advisor and committee chairperson, Dr. Xiangen Hu, for his unrelenting devotion to his students. Xiangen dedicated countless hours to helping this project through completion. I am also grateful for my other two committee members, Dr. Art Graesser and Dr. Vasile Rus. Their insightful advice significantly improved the quality of my study and of this final document. Many thanks to Jin Wang, Yuchen Liao, and Jun Xie for their assistance on this project. My gratitude goes to Abby Green, Dr. Benjamin D. Nye, Dr. Theresa Okuwumabua and Henry Hua for spending time to proofread my thesis. Finally, I would like to thank my parents, Tianhong Huang and Fengying Li. Without their support and encouragement, this would not have been possible.

ABSTRACT

Huang, Xudong. M.S. The University of Memphis. August/2013. Evaluation: Measurements of Differences between Semantic Spaces. Major Professor: Xiangen Hu, PhD.

The existing method to measure differences among semantic spaces is costly. The current study evaluates a low-cost method. Specifically, the current study uses three measurements of induced semantic structures (ISS) to measure the differences between vector-based semantic spaces. An ISS of a target word is that word's ordered nearest neighbors. Our hypothesis, which was confirmed, is that the three measurements have the ability to measure the differences between spaces. In addition, the number of nearest neighbors used by measurements has an effect on the ability. Evaluation was conducted on five Touchstone Applied Science Associates (TASA) spaces. The measured differences between spaces were compared to the objective similar pattern of TASA spaces, which follow a well-defined hierarchy. The comparison indicates that three measurements can capture the objective TASA pattern and that performance measures were better than a measurement which does not use ISS. It was concluded that the new method of measuring space differences is an apt complement to the existing method.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The last decade has seen remarkable development in vector-based semantic modeling. This technology uses real-valued vectors to represent semantics and to compute semantic relations between words in corpora. Semantic modeling starts with word co-occurrence in chosen corpora and then uses mathematical algorithms to acquire word meanings. For example, when milk and juice often occur in the same discourse environment we assume that they are semantically related. There are dozens of semantic encoding methods. Stone, Dennis, and Kwantes (2008) and Riordan and Jones (2011) reviewed 13 of them. Several popular semantic models include Hyperspace Analogue to Language (HAL; Burgess, 1998), Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), the Topic Models (Griffiths, Steyvers, & Tenenbaum, 2007) and Explicit Semantic Analysis (ESA; Gabrilovich & Markovitch, 2007).

With the fast-paced development of computer technology, generating a large number of semantic spaces in a relatively short time is achievable. The process of building semantic spaces can be summarized in three steps. People extract corpora from naturally written documents in a given domain, choose proper encoding methods, and then generate the semantic spaces for real world use. For details see Figure 1.

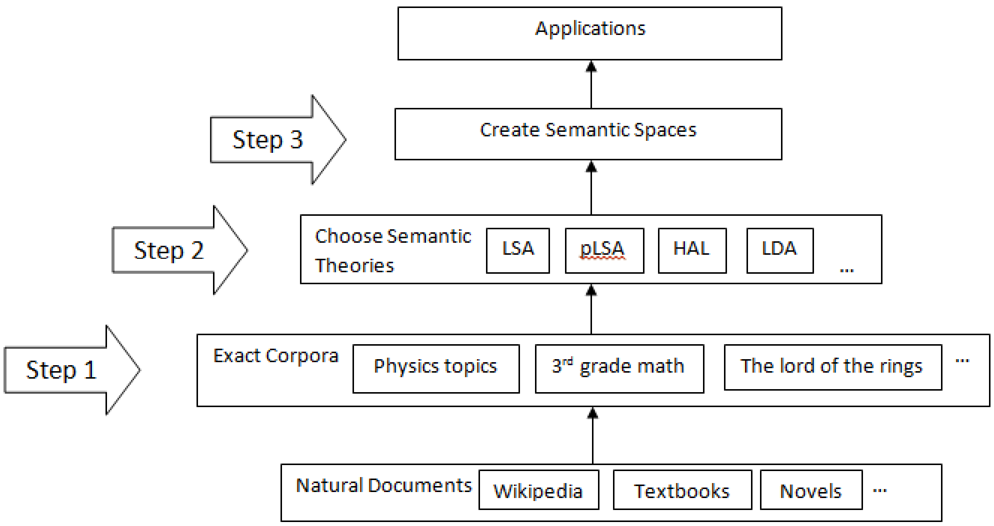

Figure 1. Three Steps of Producing Semantic Spaces

When a large number of semantic spaces occur, a new problem arises. We always need to answer the following questions when facing several space candidates: Which one is the most suitable space for a specific application? How can we compare the spaces? A correct choice significantly improves the performance of the application. In contrast, a wrong choice leads to poor performance or even non-performance. In order to choose a space accurately, scientists have developed a method to compare semantic spaces. In the past decade, many advances have been made using this method. The following chapter reviews several such studies.

CHAPTER 2

LITERATURE REVIEW

Generally, in previous studies of comparing semantic spaces, the first step is to choose a task. The task should be competent for both human and semantic spaces. The human performance collected on the task is used as the gold standard. Then semantic spaces are generated to complete the same task. If the performance of one semantic space can reach the level of human performance, but the other spaces fail, then it is argued that the successful semantic space is better than the others. In the task used by Lee, Pincombe, and Welsh (2005), the criterion of a good semantic space was the ability to emulate human judgments of similarity. These researchers (Lee et al., 2005) first built a baseline by human raters. Then they had 83 college students rate the similarity of any random pair among these 50 headline stories from Australian Broadcasting Corporation's news. An index “inter-rater correlation” was created for further evaluation. One rating for each document pair was chosen randomly and correlated with the average ratings of the remaining pairs. After 1,000 times, the average inter-rater correlation was 0.605. Once these ratings were complete, keyword, n-gram, and LSA models were chosen to compute the machine-rating similarity of the same documents. The result showed that the best LSA model had a correlation with human raters, 0.6. The best keyword and n-gram models had correlations of approximately 0.5. Other methods showed almost no correlation with the human rating.

Later researchers (Stone, Dennis, & Kwantes, 2008) extended the scope of Lee et al.’s (2005) study. Besides Lee’s (2005) corpus of news stories, Stone et al. (2008) used the Internet Movie Database (IMDB), which is a collection of celebrity gossip articles, the Touchstone Applied Science Associates (TASA) corpus, and Wikipedia. They combined the IMDB set and Lee's (2005) corpus separately with TASA and a corpus from Wikipedia to produce sub-spaces. Then, they trained six vector-based semantic models on the sub-spaces and compared their performances with human ratings. The result showed that Wikipedia performed better than TASA. Large space dimensionality increased the model similarity with human judgments. In addition, removing numbers and single letters from the corpora improved the performance of all the models. Unlike Lee et al.'s (2005) study, the vector space model had the highest judgment correlation with the human rating, 0.51.

In addition to the two above studies, Riordan and Jones (2011) used a semantic clustering task to compare the perceptual and linguistic information learned by different semantic spaces. In this study, nine semantic models were trained on the TASA corpus. Then the researchers collected their clustering performances on concrete nouns, object nouns, action verbs, and child-directed speech. As references, the researchers also used three human-generated feature models to do the same clustering tasks. The criterion of a good vector-based space was performing comparable to human-generated feature models. The result showed that several semantic spaces reached the standard, indicating that they contained sufficient semantic information that was similar to the human-generated models.

In short, the approaches used in former studies have one thing in common: Setting up a task that humans and machines can both perform. The performance of semantic spaces and humans was compared to distinguish spaces. This method does not consider the spaces' internal features, but only the input and output.

If the human performance on the task has high validity, when a semantic space meets or surpasses human performance, it is widely accepted as valid. One successful story is the Test of English as a Foreign Language (TOEFL) on LSA space (Landauer & Dumais, 1997). LSA achieved a 64.4% correct rate on 80 synonymous TOFEL questions, which is equally well as general examinees' performance (64.5%). This result has led to the popular acceptance of LSA.

However, there are some issues with this method which compares machine and human performance on the same task. First, human performance data needs to be collected for most of the tasks. Although human performance data already exists for some ready-made tasks, like the TOEFL test, data collection is costly. Second, the validity of human data varies. The national average score on TOEFL synonyms is more valid because it is coming from a larger subject sample. Word similarity rated by 20 college students is less valid because it is coming from a smaller sample. To increase the validity of human standards, researchers need to collect a large data sample, which is also time-consuming and costly. Third, in order to complete a task, specific semantic spaces need to be generated. Researchers need to select a specific corpus (e.g., child-directed speech and TASA corpus of Riordan & Jones, 2011), and train the target semantic models using the corpus to obtain a testing space. Last, when multiple spaces succeed at the same task, meaning they have all reached the human performance level, the task's power of distinguishing spaces is not sufficient. A new task for further distinction will be needed (Riordan & Jones, 2007).

If there exists a common semantic component across different semantic spaces, and the component has a numerical representation, we can use the differences within a common component to represent the space difference. Nearest neighbors of a word (known as the "target word") is such a common semantic component. Nearest neighbors are the semantically similar words to the target word in a space. At difference spaces, a word's nearest neighbors are not the same. As early as 1957, Firth indicated that "you shall know a word by the company it keeps." This view has been accepted as an important hypothesis in the research area of vector-based semantic analysis: A word's nearest neighbors represent the meaning of the target word. Therefore, using nearest neighbors could be a new method to compare semantic spaces.

The information provided by the nearest neighbors can be represented numerically (see Rationale in Chapter 3 for further explanation). Using numerical representation, the difference of nearest neighbors from several spaces indicates an ordinal ranking of the space differences. The ordinal ranking does not directly approve an absolute best space. However, we can utilize the existing well-accepted spaces and other trusted human semantic representations as references. A particular space that is minimally different from the already-evaluated spaces or semantic representations can be approved as a good space.

The new method which uses nearest neighbors is an apt complement to the current method. First, the new method can maximize the use of existing data. The references are not limited to corpus-based spaces. References can also be human semantic structure which is similar in form to the target word and its nearest neighbors. Free association norms are an excellent example here. Free association norms are human reported word association, which are widely used as a referential standard in cognitive studies. For every stimulus word, free association norms list about 10 semantic related words that people report to have thought when they first saw the word. We can view these semantic related words as nearest neighbors derived from the stimulus word. Because free association norms are human data, they are apt references for the new method. The maximum use of the existing data may reduce the need to collect new human data. Second, the new method offers large flexibility to the space candidates. The new method can use a single comparison to evaluate the spaces that differ in metric and corpus. For example, it can compare a LSA space to a probabilistic topic space. It can also compare a Wikipedia space to a LSA space of textbooks. This flexibility helps to evaluate the semantic theories/models. The third advantage is that the comparison does not produce equal results and does not need an additional task for further distinction. The new method calculates the numerical information of the nearest neighbors and reports numerical results. The results are specific to the decimal point, which can clearly separate the spaces.

Using the nearest neighbors of a target word to examine the meaning one space represents is an intuitive method. This method has been applied to some previous studies. For example, Andrews, Vigliocco, and Vinson (2009) randomly chose words in several spaces and listed their top several nearest neighbors. Different neighbors of the same target word in two spaces were used to prove that one space emphasized grounded sensory-motor senses while the other emphasized abstract encyclopedic senses.
The differences among nearest neighbors can also be used to identify words whose meanings vary across domains. For example, in order to develop a tool that can “detect semantically shifted words for translators of technical documents,” Itagaki, Aue, and Aikawa (2006) first used parsing to discover the syntactically similar words for the target

words. Then they used the overlap of the nearest neighbors as the indicator for the semantically shifted words. The less the overlap, the more one word's meaning shifted. Some researchers compare the word meanings by intuitively represent the nearest neighbors. Kievit-Kylar and Jones (2012) developed a JAVA-based tool to visualize a given word's distribution of nearest neighbors.

The previous studies mostly used nearest neighbors at the word level, either focusing on a single word's difference in meaning or using several words to illustrate the space differences. In this study, we evaluated the semantic effect of nearest neighbors at the level of complete spaces, using a straightforward evaluation to show that the information of nearest neighbors can sufficiently capture the difference between semantic spaces.

CHAPTER 3

RATIONALE

Induced Semantic Structure

Generally, in a vector-based semantic space, semantics exist at all five levels of language entities: Word, phrase, sentence, paragraph, and document. However, semantics can also be represented numerically or algebraically (for example, Turney & Pantel, 2010). Therefore, the meaning of any word can be represented by its numerical relations with other words in the same semantic space. “We call such a relation *induced semantic structure* (ISS) of the word in the given semantic space” (Hu, Cai, Graesser, & Ventura, 2005).

Induced semantic structure is the core concept of this current thesis. This concept has an origin in the field of social science. In social science, culture can be viewed as shared cognitive representation (e.g., word meaning) in human minds. Speakers of the same language share the “same” semantic structure. Romney, Boyd, Moore, Batchelder, and Brazill (1996) stated:

> The semantic structure is defined as the arrangement of the terms relative to each other as represented in a metric space in which items judged more similar are placed closer to each other than items judged as less similar. (p. 4699)

In the vector-based semantic spaces, nearest neighbors represent the meaning of a target word in the exact same way. Therefore, the concept of induced semantic structure is adopted from the field of social science and defined as the top group of ordered nearest neighbors of a word in a given semantic space.

To facilitate the understanding of induced semantic structures, an example is provided in Table 1. Table 1 lists the top 10 nearest neighbors of “hamburger” in two TASA spaces: TASA09 and TASAall. TASA spaces were produced by Touchstone Applied Science Associates, Inc. (Zeno, Ivens, Millard, & Duvvuri, 1995). The company collected reading texts from 1st grade to 1st year college students and used an encoding model called latent semantic analysis (LSA) to generate five semantic spaces. The TASA09 space used the corpus from 1st grade to 9th grade. The TASAall space used the corpus from 1st grade to 1st-year of college. LSA spaces use cosine to represent the word-to-word similarity. Basically, cosine similarity uses the cosine of the angle between two word vectors to represent whether two vectors are pointing the same direction. Value 1 means the two vectors overlap, value 0 means the vectors are perpendicular, and a value closer to 1 means the vectors are more semantically similar. In the current example, the two sets of nearest neighbors are sorted by the cosine similarity with “hamburger” in a descending order. The two ordered neighbor sets are the induced semantic structures of “hamburger” in two different contexts. Researchers manually select the number of nearest neighbors they use. We used 10 nearest neighbors in this example. The letter T is used to denote the number of nearest neighbors. So here $T = 10$.

Table 1

Two Induced Semantic Structures of 'hamburger' at Two TASA Spaces

Order	TASA09	Cosine	TASAall	Cosine
1	hamburgers	0.48	hamburgers	0.62
2	burger	0.46	macs	0.49
3	fries	0.43	fries	0.46
4	taco	0.38	chili	0.44
5	chili	0.38	steak	0.42
6	steak	0.36	menu	0.41
7	serving	0.35	burger	0.41
8	broiler	0.35	malts	0.38
9	recipe	0.34	restaurant	0.38
10	menu	0.34	cheeseburger	0.38

The concept *induced semantic structures* provides a framework that is comparable to any space with nearest neighbors, even if the two spaces do not use the same semantic encoding methods (e.g., LSA and Topic models). Furthermore, semantic spaces can be compared to semantic structure manually built by humans, such as free association norms, as long as the concept or word of the semantic structure has derived nearest neighbors. Therefore, a "best" semantic space may be identified if the space is minimally different from a human-generated semantic structure. The step of extracting induced semantic structures in the pipeline of semantic-spaces generation is shown in Figure 2.

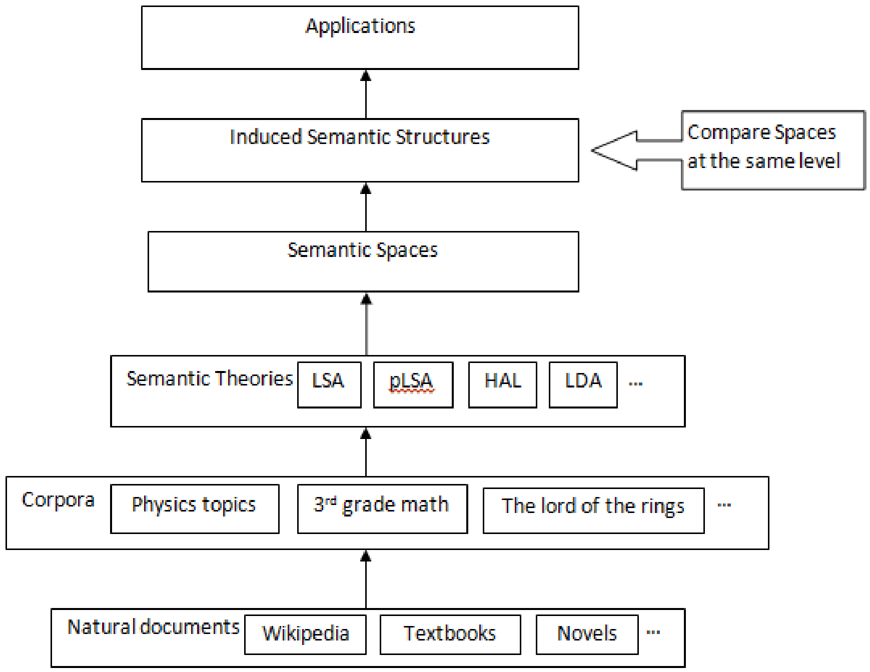

Figure 2. ISS Makes Semantic Spaces Comparable

Difference Measurements Based on Induced Semantic Structures

Before developing the measurements, Hu et al. (2005) proposed three assumptions. First, "the meaning of a word is embedded in its relations with other words." This is a well-accepted assumption in the field. Second, if a given word is shared in different semantic spaces, the relation between the semantics of the word in different spaces is "a function of the corresponding induced semantic structures." Third, the relations between any two semantic spaces are "a function of the relations of the semantic structures of all the shared words."

The second assumption emphasizes that the semantic difference of a word in two spaces can be represented by a function. In other words, we can use mathematical methods to measure the difference of nearest neighbors. Hu et al. (2005) provided three measurements, as discussed below.

Combinatorial Similarity

Combinatorial similarity calculates the overlap of two induced semantic structures. Overlap is the primary source of numerical information derived from nearest neighbors (Itagaki, Aue, & Aikawa, 2006). Suppose we have two sets of ordered nearest neighbors (ISSs) which has T nearest neighbors separately. For the two sets of top T nearest neighbors, $T \leq \min(N1, N2)$, combinatorial similarity equals the intersection of the two sets divided by the union of the two sets.

Assume a given word x, where $S1$ and $S2$ are its two sets of top T nearest neighbors in two spaces. Then the combinatorial similarity C for the word x is defined as

$$C = \frac{\|S_1 \cap S_2\|}{\|S_1 \cup S_2\|}$$

Taking Table 1 as an example, the intersection of the two sets of top 10 nearest neighbors is 5. The union is 15. Hence, the combinatorial similarity of Table 1 is 1/3. Since the combinatorial similarity uses overlap and the overlap is direct and simple, it has the widest range of applications.

Permutation Similarity

Permutation similarity considers the overlap of nearest neighbors and the order of the overlapped nearest neighbors. The positions of the overlapped words in two induced semantic structures may be different. For instance, in Table 1, “menu” places 10^{th} in one induced semantic structure but 6^{th} in the other one. So we use a permutation measurement to measure the order of the overlapped nearest neighbors. We call the measured value the permutation value. Thereby, the permutation similarity of two induced semantic structures is the product of its combinatorial similarity and its permutation value. In the current thesis, we use Spearman’s rank correlation (Spearman, 1904) as the permutation measurement. The algorithm of Spearman's rank correlation will be introduced in the next chapter.

Assume the permutation value is p. Then the permutation similarity P is defined as

$$P = pC$$

Permutation similarity extracts more information from induced semantic structures than combinatorial similarity. The adding of ordinal information provides more preciseness to the measurement of space difference.

Quantitative Similarity

Quantitative similarity also measures both the overlap of the nearest neighbors and the order of the overlapped nearest neighbors. The order here is from the nearest neighbors' quantitative similarity value to the target word. When two spaces are built from the same semantic model, they are in the same metric. For example, two LSA spaces use cosine to represent word similarity. When we compare two such LSA spaces, the order information can be obtained by directly calculating the order of the cosine instead of the order of nearest neighbor words. For instance, the overlapped nearest neighbors in Table 1 are "hamburgers," "burger," "fries," "chili," "steak," and "menu." In space TASA09, the cosine values of "hamburger" and its nearest neighbors are 0.48, 0.46, 0.43, 0.38, 0.36, and 0.34. In space TASA12, the cosine values are 0.62, 0.41, 0.46, 0.44, 0.42, and 0.41. Then, Pearson's correlation (Pearson, 1907) of the cosine values can measure the order of two induced semantic structures.

For two induced semantic structures, the quantitative similarity is their combinatorial similarity multiplied by Pearson's correlation of their cosine values.

Assuming Pearson's correlation of the quantities is q, the quantitative similarity Q is defined as

$$Q = qC$$

It is worth emphasizing that, when comparing spaces from different models, we use the permutation similarity. For example, LSA spaces use cosine to represent similarity. Topic models use KL divergence to represent similarity. LSA models and Topic models are not in the same metric. In this case, we would compare the order of the nearest neighbors in the LSA model space to the nearest neighbors in the Topic model space, regardless of their similarity values. By doing this we use the order place of the neighbors to calculate the order information.

A Competing Measurement: Correlation of the Cosine

The goal of the current paper is to empirically evaluate the use of induced semantic structures in evaluating semantic spaces. In other words, we use ordered nearest neighbors to measure the difference of spaces. As experiments have a control group, we also have a competing measurement which does not use nearest neighbors to measure space differences. Here we develop a measurement called the correlation of the cosine. The algorithm is the correlation value of random words' cosine matrix. The cosine value is an excellent indicator of word-to-word similarity in a single space. Though cosine similarity cannot directly measure the similarity of words across spaces, it is easy to think about bridging two spaces using the correlation of cosine similarity. Therefore, we use Pearson's correlation of the cosine as our competing measurement. First, we randomly select a large number of words as a sample from all common words across spaces. Then we obtain the cosine similarity for every pair of sample words in single spaces. This cosine similarity can be represented as a cosine matrix whose rows and columns are the sample words. The matrix has an equal number of rows and columns. Then we compute Pearson’s correlation of the cosine matrices of two spaces. The correlation value is the measure of space difference.

CHAPTER 4

METHOD

Design

The Hu method (Hu et al., 2005) measures the difference between semantic spaces. One direct evaluation of the measurements is to find several semantic spaces with objective difference pattern and compare the result of the Hu method with this pattern. The TASA spaces are the spaces which have objective difference pattern.

TASA spaces were produced with reading texts from 1st grade to 1st year college. Space TASA03 includes texts from 1st grade to 3rd grade. TASA06 includes texts from 1st grade to 6th grade. Following the same pattern, TASA09 contains all texts from 1st grade to 9th grade. TASA12 contains all texts from 1st grade to 12th grade. TASAall contains all texts from 1st grade to 1st year of college. Since the spaces are added, the neighbor spaces should have a higher similarity than the others. For example, TASA03 should be more similar to TASA06, than to TASA09, TASA12, and TASAall. We use this pattern as the reference to evaluate the four space similarity measurements. The reference pattern is shown in Table 2.

Table 2

Similarity Pattern of TASA Spaces

Space	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03	identical	high	low	lower	lowest
TASA06	high	identical	high	low	lower
TASA09	low	high	identical	high	low
TASA12	lower	low	high	identical	high
TASAall	lowest	lower	low	high	identical

We used the four measurements introduced at the last chapter to measure the difference of semantic spaces pairwise. By comparing against the reference pattern in Table 2, we can know which measurement(s) catches the objective pattern.

Materials: TASA spaces

TASA spaces were generated by Touchstone Applied Science Associates, Inc. to develop The Educator's Word Frequency Guide (Zeno et al., 1995). After generation, they have been widely used, generally well accepted, used in various research projects and applications (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Riordan & Jones, 2011). The specifics of the spaces are listed in Table 3. As shown in the table, the TASA spaces are added. For example, the corpus of the space TASA06 contained the 6,974 documents from TASA03, and added the other 10,975 documents from the 4th to 6th grade reading. The number of the added documents varies for difference spaces. Because the TASA space of 1st year college included the entire documents from 1st grade to 1st year of college, it is referred to as TASAall space at the current thesis.

Table 3

Specifics of Five TASA Spaces

Space	Grade	Number of Documents	Document Added	Number of Terms	Terms Added	Number of Dimensions
TASA03	3	6,974	——	29,315	——	432
TASA06	6	17,949	10,975	55,105	25,790	412
TASA09	9	22,211	4,262	63,582	8,477	407
TASA12	12	28,882	6,671	76,132	12,550	412
TASAall	college	37,651	8,769	92,409	16,277	419

Manipulation

Combinatorial Similarity. The process began with random selection of 1,000 common words among spaces. Since TASA03 is covered by all the other four spaces, a random selection of 1,000 words was chosen from TASA03 as the sample. For each word at each space, we obtained its top 50, 100, and 200 nearest neighbors and computed the combinatorial similarity. The average of the 1,000 sample words' similarity was used to calculate the reported results for this measurement.

Permutation Similarity. This step used the same 1,000 words and the same 50, 100, and 200 nearest neighbors to compute permutation similarity. In the current thesis, Spearman's rank correlation (Spearman, 1904) was used to get the permutation difference. Spearman's rank is designed for ordinal values. It fulfills our need to calculate the correlation of two sets of ordered nearest neighbors. The detailed steps are as follows. First, we ordered the *n* overlapped nearest neighbors of one target word across two spaces from 1 to *n* in order of the largest cosine. Then we computed the Spearman's rank correlation for the two sets. For instance, in Table 1, the overlapped nearest neighbors are "hamburgers," "burger," "fries," "chili," "steak," and "menu." Their order in space TASA09 is 1, 2, 3, 4, 5, and 6; while their order in space TASAall is 1, 6, 2, 3, 4, and 5. Their Spearman's rank correlation $r = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 20}{6(6^2-1)} = 0.429$. As mentioned in the rationale chapter, the permutation similarity is the multiplication of the permutation value and the corresponding combinatorial similarity. Hence, the permutation similarity in the current thesis is the product of Spearman's rank correlation and the corresponding combinatorial similarity. We use the average of the permutation similarity values of the 1,000 words as the reporting result for this measurement.

Quantitative Similarity. The measurement of the quantitative similarity also uses the same 1,000 words and their 50, 100, and 200 nearest neighbors. As mentioned in the rationale chapter, Pearson's correlation was adopted to obtain the correlation of the neighbors' cosine values. Pearson's correlation is well-described in Spearman (1907), so we will not go into the details here. According to the discussion of the rationale chapter, the quantitative similarity of two spaces is the multiplication of the quantitative difference and the corresponding combinatorial similarity. So we use the product of Pearson's correlation and the corresponding combinatorial similarity as our measurement. The reporting result for this measurement is the averages of the quantitative similarity values of the 1,000 sample words.

The Competing Measurement: Correlation of the Cosine. The competing measurement is the correlation of the cosine which does not include the information of nearest neighbors. It only considers the words themselves and their cosine similarity. This measurement used the same selection of 1,000 words. In each space, the cosine values of the 1,000 words were obtained. Then, Pearson's correlation was applied to every pair of semantic spaces to get the correlation of the 1,000 words' cosine values across two spaces. Pearson's correlation values are the reporting values of the measurement.

Evaluation the Performance of the Measurements

Since we have four measurements trying to catch the difference pattern of TASA spaces, it is necessary to compare the performance of these four measurements. When a measurement has the ability to measure the space difference, it should report a large similarity value for the should-be-high space pair, e.g., TASA03 and TASA06. Also, it should report a small value for the should-be-low space pair, e.g., TASA03 and TASA12.

When we calculate the average of all should-be-high space pairs and the average of all should-be-low space pairs, the former should be larger than the latter. Therefore, if we divide the former by the latter, the quotient must be greater than 1. For example, in Table 2, the cells TASA03-TASA06, TASA06-TASA09, TASA09-TASA12, and TASA12-TASAall have the highest similarity values. The rest cells have lower similarity values. Then the average of the four cells (TASA03-TASA06, TASA06-TASA09, TASA09-TASA12, and TASA12-TASAall) divided by the average of the rest of the cells must be larger than 1. We call this the ratio of performance. Following the same algorithm, we have two other ratios of performance. One is the ratio of second-highest average (TASA03-TASA09, TASA06-TASA12 and TASA09-TASAall) to the average of lower remaining cells (TASA03-TASA12, TASA06-TASAall and TASA03-TASAall). The other one is the ratio of the third-highest average (TASA03-TASA12, TASA06-TASAall) to the lower remaining cell (TASA03-TASAall). When several measurements all have a ratio of performance larger than 1, the measurements with the largest ratio is the best measurement.

CHAPTER 4

RESULTS

Combinatorial Similarity

The combinatorial similarities of TASA spaces using 50, 100, or 200 neighbors are in Table 4. Comparing the result table with Table 2, the reference pattern, we observe that the patterns match substantially at all three levels of nearest neighbors. Most should-be-high values are large, and all the should-be-low values are small. The result generally indicates that the neighboring spaces have higher similarity than the not-neighboring spaces. The only exception is the similarity of TASA06 and TASA03. Its similarity was considered to be higher than the TASA6-TASA12 combination and the TASA6-TASAall combination. However, the actual TASA06-TASA03 similarity is lower than the TASA6-TASA12 combination and the TASA6-TASAall combination.

Checking Table 3 gives us a clue as to why the space TASA03 is odd. The corpus of TASA03 had 6,974 documents and the TASA06 had 17,949 documents. Because TASA06 is an added space from TASA03, we know that TASA03 corpus only composes 38.85% of TASA06 corpus. In contrast, TASA09 has 22,211 documents and TASA12 has 28,882 documents. TASA06 composes 80.81% of TASA09 and composes 62.15% of TASA12. Hence, the proportion of the overlapped corpus between TASA03 and TASA06 is much lower than the ones of TASA06-TASA09 and TASA06-TASA12. When we argued that the neighboring spaces have higher similarity, we assumed that the documents added to create the higher grade spaces were in the same proportion. However, we neglected to consider that the quantity of the added documents changes dramatically in TASA spaces. Therefore, the previous reference pattern was not precise.

Rather, the order of document overlap percentage is a more accurate reference. Higher document/term overlap indicates a higher space similarity. We calculated the overlap percentages of the documents and terms in Table 5. The third column of Table 5 shows that TASA06 compared to TASA03 has a higher document/term overlap with TASA09, TASA12, and TASAall. Therefore, the space most similar to TASA06 is TASA09, the second similar space is TASA12, the third similar space is TASAall, and the least similar space is TASA03. Using document/term overlap as the reference for the TASA similarity pattern is better than the original reference which only used neighbors or not to indicate the similarity relations. Neighbor or not cannot distinguish a target space's relation between the left and right neighbors. Also, the neighboring TASA spaces are not always most similar to each other. From now on, we will use the document/term overlap as the reference of space similarity. Since document and term overlap have the same pattern, we will only use document overlap as the reference for the following comparisons.

A direct observation of Table 4 and Table 5 shows that the order of the measured pattern matches the order of the reference pattern. We use the pattern of TASA03 as an example. Please read the tables by columns. Table 5 indicates that the most similar space of TASA03 is TASA06, the second similar space is TASA09, the third similar space is TASA12, and the least similar space is TASAall. In Table 4, for the condition of 50, 100, and 200 neighbors, the most similar space of TASA03 is also TASA06. The second similar space is TASA09. The third similar space is TASA12, and the least similar space is TASAall. Checking all columns shows that for every space, the order of measured similarity with other spaces (most similar, second similar, third similar and least similar) matches the order of the reference pattern.

Table 4

Combinatorial Similarity of 50, 100, and 200 Nearest Neighbors

	TASA03	TASA06	TASA09	TASA12	TASAall
			50 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.344214254	0.266635296	0.204004956	0.140846895
TASA06	0.344214254		0.793473852	0.589087804	0.40292577
TASA09	0.266635296	0.793473852		0.757795556	0.517648809
TASA12	0.204004956	0.589087804	0.757795556		0.718799602
TASAall	0.140846895	0.40292577	0.517648809	0.718799602	
			100 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.239618594	0.184218942	0.140596692	0.09743117
TASA06	0.239618594		0.678051602	0.466014867	0.299184047
TASA09	0.184218942	0.678051602		0.640037161	0.400126861
TASA12	0.140596692	0.466014867	0.640037161		0.601549747
TASAall	0.09743117	0.299184047	0.400126861	0.601549747	
			200 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.154735509	0.119968961	0.092592508	0.064368022
TASA06	0.154735509		0.524083227	0.333765482	0.203276944
TASA09	0.119968961	0.524083227		0.490895529	0.279785651
TASA12	0.092592508	0.333765482	0.490895529		0.450031315
TASAall	0.064368022	0.203276944	0.279785651	0.450031315	

Table 5

Document and Term Overlap of the TASA spaces

			Document Overlap (%)		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		38.85	31.40	24.15	18.52
TASA06	38.85		80.81	62.15	47.67
TASA09	31.40	80.81		76.90	58.99
TASA12	24.15	62.15	76.90		76.71
TASAall	18.52	47.67	58.99	76.71	

(Continued)

Table 5

Document and Term Overlap of the TASA spaces

	Term Overlap (%)				
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		53.20	46.11	38.51	31.72
TASA06	53.20		86.67	72.38	59.63
TASA09	46.11	86.67		83.52	68.80
TASA12	38.51	72.38	83.52		82.39
TASAall	31.72	59.63	68.80	82.39	

Permutation Similarity

The permutation similarity of 50, 100, and 200 neighbors are in Table 6. A direct observation of Table 6 and Table 4 also indicates that the order of the measured permutation similarity with other spaces (most similar, second similar, third similar and least similar) matches the order of the reference pattern.

Table 6

Permutation Similarity of 50, 100, and 200 Nearest Neighbors

	50 Neighbors				
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.152559166	0.103861707	0.069525177	0.041892245
TASA06	0.152559166		0.528877731	0.320152052	0.187987667
TASA09	0.103861707	0.528877731		0.479286204	0.268592280
TASA12	0.069525177	0.320152052	0.479286204		0.453724493
TASAall	0.041892245	0.187987667	0.268592280	0.453724493	
	100 Neighbors				
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.107631405	0.073214435	0.049100146	0.030094322
TASA06	0.107631405		0.454366927	0.254989610	0.140226235
TASA09	0.073214435	0.454366927		0.408639143	0.210215822
TASA12	0.049100146	0.254989610	0.408639143		0.380826021
TASAall	0.030094322	0.140226235	0.210215822	0.380826021	

(Continued)

Table 6

Permutation Similarity of 50, 100, and 200 Nearest Neighbors

			200 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.068843650	0.048316261	0.032257387	0.019656631
TASA06	0.068843650		0.351138703	0.182090137	0.095180293
TASA09	0.048316261	0.351138703		0.317028755	0.149933310
TASA12	0.032257387	0.182090137	0.317028755		0.288548677
TASAall	0.019656631	0.095180293	0.149933310	0.288548677	

Quantitative Similarity

The permutation similarity of 50, 100, and 200 neighbors are in Table 7. A direct observation of Table 7 and Table 4 also indicates that the order of the measured quantitative similarity with other spaces (most similar, second similar, third similar and least similar) matches the order of the reference pattern.

Table 7

Quantitative Similarity of 50, 100, and 200 Nearest Neighbors

			50 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.178309050	0.122493035	0.083086320	0.050401536
TASA06	0.178309050		0.598365601	0.372594232	0.218953844
TASA09	0.122493035	0.598365601		0.545325379	0.310055483
TASA12	0.083086320	0.372594232	0.545325379		0.515308216
TASAall	0.050401536	0.218953844	0.310055483	0.515308216	
			100 Neighbors		
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.128495256	0.087365445	0.058752303	0.035786705
TASA06	0.128495256		0.525464474	0.305021294	0.167745966
TASA09	0.087365445	0.525464474		0.475230983	0.249503040
TASA12	0.058752303	0.305021294	0.475230983		0.442757058
TASAall	0.035786705	0.167745966	0.249503040	0.442757058	

(Continued)

Table 7

Quantitative Similarity of 50, 100, and 200 Nearest Neighbors

	200 Neighbors				
	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.084040040	0.058373183	0.039467686	0.024012791
TASA06	0.084040040		0.414300504	0.222927910	0.116384222
TASA09	0.058373183	0.414300504		0.375322648	0.180654797
TASA12	0.039467686	0.222927910	0.375322648		0.339709983
TASAall	0.024012791	0.116384222	0.180654797	0.339709983	

Correlation of the Cosine

The correlation of the cosine between TASA spaces are in Table 8. A direct observation of Table 8 and Table 4 also indicates that the order of the measured similarity with other spaces (most similar, second similar, third similar and least similar) matches the order of the reference pattern.

Table 8

Correlation of the Cosine

	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		0.340395010	0.284931136	0.237597438	0.20037843
TASA06	0.340395010		0.765601621	0.594569012	0.46166667
TASA09	0.284931136	0.765601621		0.743623101	0.56225027
TASA12	0.237597438	0.594569012	0.743623101		0.73265338
TASAall	0.200378427	0.461666670	0.562250266	0.73265338	

Ratio of Performance on TASA Spaces

Because all four measurements can extract the pattern of the TASA spaces, comparison of the performance was conducted to distinguish the measurements. As mentioned in the rationale chapter, we used the ratio of performance to indicate the level of performance. A larger value represents a better performance. The algorithm of the ratio examines the multiple of the should-be-high averages to the should-be-low averages. The algorithm does not change. But because the reference pattern has been updated, we will also update the information of the should-be-high cells in Table 9 and the should-be-low cells. We simplified the reference pattern of the document overlap of Table 5 to a similarity ranking in Table 9. Please read the table by columns. In the table, 1 means the most similar, 4 means the least similar. Hence, the three kinds of ratios are: The average of 1 divided by the average of 2, 3, and 4; the average of 2 divided by the average of 3 and 4; the average of 3 divided by the average of 4. The values of the ratios were calculated in Table 10. The result indicates that the three measurements with nearest neighbors perform better than the one without nearest neighbors, the correlation of the cosine. Within the three measurements with nearest neighbors, the permutation and the quantitative similarities perform better than the combinatorial similarity. In addition, the increase of the number of neighbors increases the measurement performance.

Table 9

TASA Spaces Similarity ranked by Document Overlap

Space	TASA03	TASA06	TASA09	TASA12	TASAall
TASA03		4	4	4	4
TASA06	1		1	3	3
TASA09	2	1		1	2
TASA12	3	2	2		1
TASAall	4	3	3	2	

Table 10

Ratio of Performance on TASA Spaces

	Neighbors	1/(2+3+4)	2/(3+4)	3/4
Combinatorial	50	1.686143568	1.773944379	1.930232457
	100	1.828025153	1.938711379	2.11393885
	200	1.958921333	2.08169968	2.24319244
Permutation	50	2.09473302	2.251585108	2.524207351
	100	2.24533377	2.447956484	2.739273311
	200	2.388319814	2.652553958	2.938800296
Quantitative	50	2.055934781	2.210196585	2.483319352
	100	2.204064885	2.409161463	2.740629125
	200	2.344547252	2.59899659	2.939537888
Correlation of the Cosine	Not applicable	1.545301953	1.629531496	1.83412672

CHAPTER 5

DISCUSSION

The purpose of this study was to evaluate a new method of measuring the differences between semantic spaces. The new method has large flexibility and is an apt complement to the current method which uses the human tasks as criteria. By using the common semantic component across spaces, the nearest neighbors of the words, the new method maximize the use of the existing data and can work on semantic spaces from different encoding methods and corpora. The difference pattern of five TASA spaces was used to test the ability of the method. The result suggests that the method works efficiently.

The TASA spaces were added spaces. In other words, the corpora of the lower grades were included in the corpora of the higher grades. Therefore, the overlap of the corpora created an objective similarity pattern between TASA spaces. Intuitively, we thought that the neighboring spaces would have higher similarity than the non-neighboring spaces. That was the original reference pattern of the current study. However, this judgment had a hidden precondition: The number of the documents added to the previous corpora should generally have the same proportion. That was not completely true for the TASA spaces. The TASA06 space contains 17,949 documents, which is 2.6 times that of TASA03. The corpus of TASA03 only composed 38.85% of TASA06. But TASA06 composed 80.81% of TASA09 and 62.15% of TASA12. Therefore, TASA06 is obviously closer with TASA09 and TASA12 than TASA03. In order to reflect the precise similar pattern of the TASA spaces, we use the document overlap percentage of the spaces to express the similar pattern. Larger overlap means more similar spaces. Details are in Table 5. The pattern is almost the same as the original reference pattern except that the two most similar spaces of TASA06 are TASA09 and TASA12, instead of TASA03 and TASA09.

According to the results, the method extracts the TASA pattern precisely. The direct comparison of the measured similarity and the reference pattern show a matched order. For all four measurements, the decreasing order of every TASA space with all other spaces are the same as that reference order, the order of document overlap percentages. Hence, the similarity measurement measures the real pattern of the spaces. In addition, though the values extracted from the measurements are generally ordinal, the difference between two values reflects internal information to a certain extent. For example, TASA12 has almost equal similarity with TASA09 and TASAall because the document overlap percentages of TASA12-TASA09 and TASA12-TASAall are the same down to two decimal places. Correspondingly, the similarity values of the pair TASA12-TASA09 and TASA12-TASAall are much closer compared to the values of the other pairs.

The direct observation of the result tables provides a basic knowledge of the measurements' ability. We further used a ratio of should-be-high values to should-be-low values to distinguish the performance of four measurements. The result shows that the three measures using nearest neighbors perform better than the correlation of the cosine, which does not consider the information of nearest neighbors. It infers that the nearest neighbors provide more information for a target word than the target word's own similarity with other random words. The sufficient semantic information contained in nearest neighbors has been proved by multiple studies. Widdows (2003) used the unknown words' nearest neighbors to automatically classify the meaning of the words and therefore map the unknown words into taxonomy. Jones and Mewhort (2007) and Andrew et al. (2009) used sample words' nearest neighbors to distinguish the semantic emphases of specific models. The current study proves once again that the method of nearest neighbor is valid and efficient.

Within the three measurements of nearest neighbors, the permutation and the quantitative similarity perform better than the combinatorial similarity. Obviously, in addition to the number of overlap in nearest neighbors, permutation and quantitative similarities contain the order information of the nearest neighbors. Permutation has the order of nearest neighbors. The quantitative similarity has the order of the similarity values to the target word. The adding of the order information helps distinguish semantic spaces. One thing to address is that the methods to get permutation and quantitative difference affect the pattern extracting ability. Different methods generate different performance. A strong method improves the ability. In the current study, we used Spearman's rank correlation and Pearson's correlation. Since Spearman's rank correlation is the variation of Pearson's correlation which keeps most of the information of Pearson's correlation, the levels of performances of the permutation and the quantitative similarity are very close.

Three numbers of nearest neighbors were considered in the current study. Macroscopic observation of the similarity tables did not show a significant difference between the results of 50, 100, and 200 nearest neighbors. However, the ratio of the performance indicates that the pattern extracting ability increases slightly when the number of nearest neighbors increases. Widdows (2003) also reported that the number of the nearest neighbors affected the performance of the classification. It should be noted that, the current finding is the initial application of the nearest neighbors on space difference measurement. We remain cautious about the finding. The impact of the number of nearest neighbors on the three similarity measurement is a curve with multiple turning points. The points 50, 100, and 200 are three samples from the curve, which only offer a glimpse of the complete phenomenon. The turning points may occur after 50, 100, or 200 neighbors. In Figure 3 there is a sample curve of the combinatorial similarities of the word “hamburger” with nearest neighbors from 1 to 500. The permutation and quantitative similarities of “hamburger” with nearest neighbors from 1 to 160 are in Figure 4. For the combinatorial similarity, the curve goes smoothly from 50 nearest neighbors. For the quantitative similarity, turning points occur at 15, 50, 60, and 100 nearest neighbors. Therefore, if we want to have a comprehensive understanding on the effect of the numbers of nearest neighbors, a study of the whole curve is needed. That is one further direction of the current study.

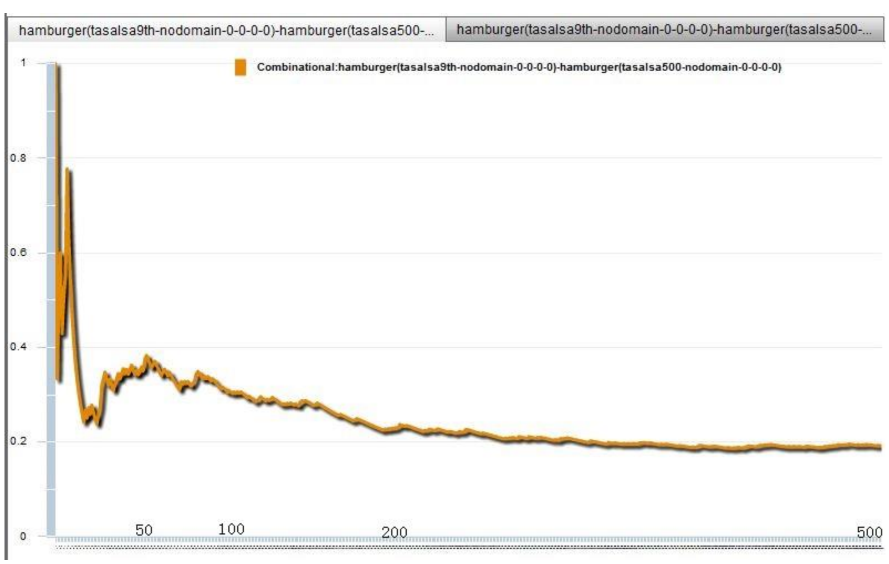

Figure 3. Combinatorial similarities of ‘hamburger’ with nearest neighbors from 1 to 500

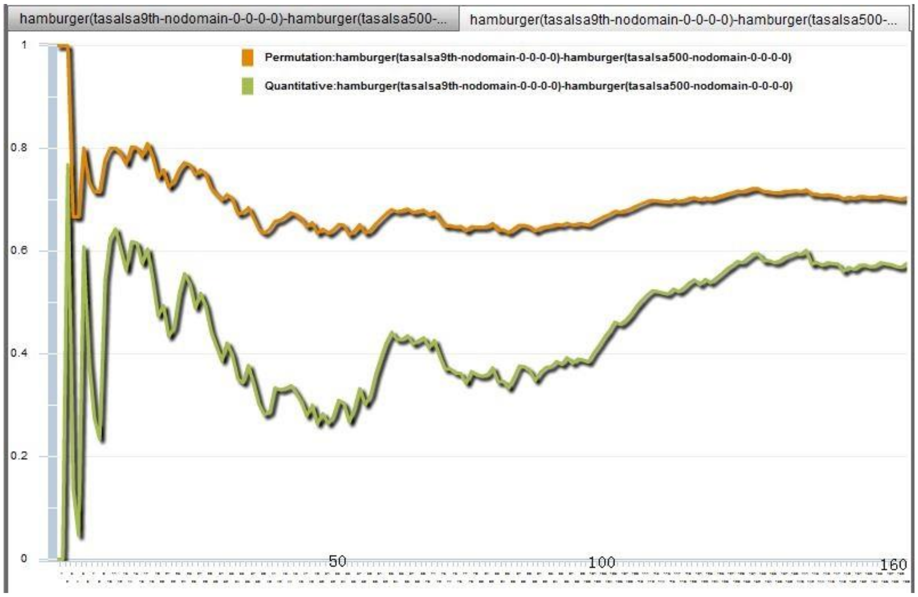

Figure 4. Permutation and quantitative similarities of 'hamburger' with nearest neighbors from 1 to 160

Limitations and Future Directions

The present study has two major limitations that need to be addressed in future studies. First, as mentioned above, the sample numbers of the nearest neighbors are limited. As a first attempt, the present study proved that the number of nearest neighbors has an effect on the ability of the measurements. A comprehensive examination of effect will provide an accurate description of the impact. For example, with the increase in the number of nearest neighbors, the similarity value changes dramatically at the first part and then goes smooth. Therefore, finding the complete impact trend and detecting where to stop adding nearest neighbors will be interesting questions.

The second limitation pertains to the type of the semantic theories considered in the study. LSA is a popular theory. But in addition to it, pLSA, the topic models, and many other spaces are also widely accepted. The impact of the nearest neighbors may vary among theories. The evaluation of other semantic theories is needed to complement our findings on the LSA spaces.

In order to understand the algorithm more comprehensively, further study may also consider different word types. Widdows (2003) reported that classification using nearest neighbors is obviously better for common nouns than for verbs. In the current study we sampled random words from the corpus which contained different word types, e.g., nouns, verbs, adjectives, and adverbs. If we separated the words by type and compared their results on similarity measurements, we may find that different types of words have different abilities.

Implications

The method of induced semantic structure evaluates the difference between semantic spaces using the information of nearest neighbors. Nearest neighbors are common semantic components among vector spaces. Hence, this method can be applied to a very wide field. The method helps reduce the cost of collecting human data for space evaluation.

In addition, this approach is an application of the nearest neighbors. Hu et al. (2005) is the initial theory to use neighbors to measure the difference between spaces. This successful evaluation supports the theory and indicates the further directions.

CHAPTER 6

CONCLUSION

The current study verifies that the method of nearest neighbors works effectively in measuring differences between semantic spaces. Using the nearest neighbors of the target words to extract space difference is more efficient than directly using the relation between target words themselves. The number of nearest neighbors has an effect on the ability to measure space difference. A comprehensive understanding of the effect needs more exploration.

REFERENCES

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*, 463–498. doi:10.1037/a0016261

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, and Computers, 30*(2), 188-198. doi: 10.3758/BF03200643

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-Based explicit semantic analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (Vol. *7*, pp.1606-1611). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244. doi: 10.1037/0033-295X.114.2.211

Hu, X., Cai, Z., Graesser, A. C., & Ventura, M. (2005). Similarity between semantic spaces. *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp.995-1000). http://csjarchive.cogsci.rpi.edu/

Itagaki, M., Aue, A., & Aikawa, T. (2006). Detecting inter-domain semantic shift using syntactic similarity. *Proceedings of the 5th International Conference on Language Resources and Evaluation*(pp.2399-2402). http://www.lrec-conf.org/proceedings/lrec2006/

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1-37. doi:10.1037/0033-295X.114.1.1

Kievit-Kylar, B., & Jones, M. N. (2012). Visualizing multiple word similarity measures. *Behavior research methods, 44*(3), 656-674.

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*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Lee, M., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*(pp.1254-1259). http://csjarchive.cogsci.rpi.edu/

Pearson, K. (1907). *On further methods of determining correlation* (Vol. *4*). Fetter Lane, E. C.: the Cambridge University Press.

Riordan, B., & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 599-604). http://csjarchive.cogsci.rpi.edu/

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3(2)*, 303-345. doi: 10.1111/j.1756-8765.2010.01111.x

Romney, A. K., Boyd, J. P., Moore, C. C., Batchelder, W. H., & Brazill, T. J. (1996). Culture as shared cognitive representations. *Proceedings of the National Academy of Sciences*, *93*(10), 4699-4705.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, *15*(1), 72-101.

Stone, B. P., Dennis, S. J., & Kwantes, P. J. (2008). A systematic comparison of semantic models on human similarity rating data: The effectiveness of subspacing. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp.1813-1818). http://csjarchive.cogsci.rpi.edu/

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141-188.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 197-204). Association for Computational Linguistics.

Zeno, S. M., Ivens, S. H., Millard, R. T. & Duvvuri, R. (1995). *The educator's word frequency guide*. Breward, NY: Touchstone Applied Science.