Assessing the Effectiveness of Corpus-based Methods in Solving SAT Sentence Completion Questions

Eugene Tang
Supervised by Professor Christiane Fellbaum

2015-04-30

Abstract

This paper studies different corpus-based algorithms through which to answer SAT sentence completion questions. SAT sentence completion questions assess how well different words fit into a sentence, and the ability to answer such types of questions have wide implications in optical character and speech recognition post-processing as well as word-suggestion programs. In our study, we analyze several statistical corpus-based methods through which to answer such questions, including normalized pointwise information, co-occurrence frequencies, latent semantic analysis, and the word2vec neural net implementations of continuous bag of words (CBOW) and continuous skip-gram (CSKIP) models. We find that the co-occurrence frequency method by itself has a near state-of-the-art performance with 52% correctness and that combining the co-occurrence frequency method with CBOW and CSKIP results in a state-of-the-art performance of 59%. The results of this study demonstrate that local context is a fairly strong measure in determining how well a word fits in a sentence and that exploration of non-similarity based methods may be required to further increase the ability of computers to answer such questions.

1. Introduction

Gap-filling questions are a class of questions in which a sentence is provided with one or more words replaced with a gap. The participant must then select the best word to fill each gap from a set of possible choices. In this paper, we study a specific type of gap-filling question—the sentence completion questions from the Scholastic Aptitude Test (SAT), a standardized exam used for college admissions. Each SAT sentence completion question contains a sentence containing one or two blanks, and five answer choices containing a single word (or phrase) or a pair of words (or phrases) respectively. The goal of the test-taker is to then select the answer choice that best fills the blank(s) in the sentence. Two example questions are shown in Figure 1.

> The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including ------- ones.
>
> (A) outmoded (B) figurative (C) experimental
> (D) cursory (E) permanent
>
> Cookery ------- the ------- of science, for the observations of prehistoric cooks laid the foundations of early chemistry.
>
> (A) ignored . . precision
> (B) advanced . . development
> (C) retarded . . supremacy
> (D) aided . . decline
> (E) betrayed . . methodology

Figure 1: Two example SAT sentence completion questions [4]. The correct answers are C and B respectively.

The questions present a unique challenge since the answer choices are often closely related to each other, differing only in connotation. Furthermore, the questions generally test understanding of words that do not appear in everyday speech (see Figure 2), and some questions contain two blanks, in which other factors such as the relation between the words in the blanks must also be taken into account. However, since the questions are designed to assess knowledge of English, each question contains all the information necessary to be answerable without any other context. Thus the sentence completion questions are a unique class of questions through which to evaluate different language processing methods.

In addition to evaluating different language processing methods, developing methods through which to answer these questions thus have several possible applications. For example, one obstacle

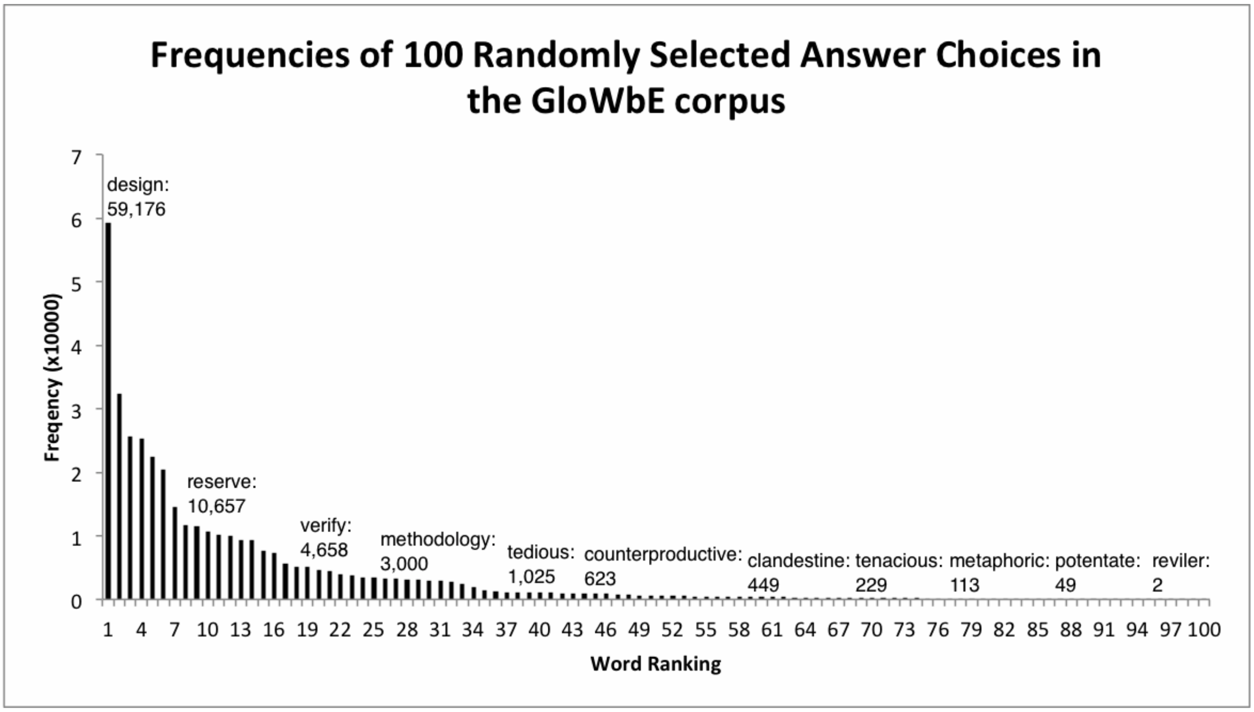

Figure 2: Graph of word frequencies of 100 randomly selected answer choices.

often encountered in optical character recognition (OCR) and speech recognition is that the translation from raw input to text is imperfect. A post-processing step is thus needed to "clean" the raw output at a lexical or semantic level. If a word is hard to recognize in the raw input, some OCR programs provide a list of possible candidates for the word [10, 22]. A gap-filling technique could be used to determine which of the candidate choices is most likely. Another possible application is to assess the quality of sentence completion questions. Since sentence completion questions are often used to assess cognitive ability, such as in the SATs, the quality of the questions are of paramount importance. Different methods to solve such questions could thus be used to assess factors such as the quality of the distractors (the incorrect answer choices) or if the sentence contains enough information for the question to be answered without additional context. Other possible applications of such methods also include finding search results that best fit a query or giving word suggestions when writing documents (e.g. using *mob* instead of *group*).

Currently, the state-of-the-art results for solving SAT sentence completion questions report 53% correctness using a combination of latent semantic analysis (LSA) and a Good Turing language model, and the state-of-the-art results for an individual method is 46% using LSA [25]. In this study, we tested several corpus-based methods that have not before been evaluated on SAT sentence completion questions, including co-occurrence frequencies, pointwise mutual information (PMI), and the word2vec[1] implementations of continuous skip-gram (CSKIP) and continuous bag-of-words (CBOW). Beyond testing these methods, we also take a novel approach on answering the questions by creating a solver that not only decides which answer choice is correct but also evaluates how well each answer choice fits in the sentence, thus discerning which answer choices are the second, third, fourth and fifth best choices as well.

Through this study, we found co-occurrence frequencies to be the best individual performer, with 52% correctness. Furthermore, our final solver has a state-of-the-art performance of 59% by combining the co-occurrence frequency, CSKIP, and CBOW models.

2. Data

In this study we used three main sources of data. To evaluate the performance of the different methods, we assessed the methods on 108 SAT sentence completion questions obtained from official SAT practice exams between 2003 and 2014 [4, 5, 6, 7, 8, 9] and their corresponding answer keys. Statistics regarding the number of blanks and the relative difficulties[2] of the SAT sentence completion questions used are shown in tables 1 and 2. To train the methods, we used the GloWbE corpus of English websites.[3] The entire corpus contains 1.8 billion words. However, since the SATs are mainly used for college admissions in the United States, we only use the subset of American websites in GloWbE, reducing the effective corpus size to ~300 million words.

Number of Blanks	Number of Questions
One blank	60
Two blanks	48

Table 1: Distribution of number of blanks.

[1] https://code.google.com/p/word2vec/

[2] In the answer keys of each of their practice exams, the CollegeBoard provides a difficulty measure for each of the questions. The difficulty ranges from 1 to 5, with 1 being the easiest and 5 being the hardest. For certain years, the difficulty was measured in terms of "E", "M", an "H" instead of the usual 1-5 scale. For those years, "E" was converted to 1, "M" was converted to 3, and "H" was converted to 5.

[3] http://corpus.byu.edu/glowbe/

Difficulty	Number of Questions
1	21
2	18
3	30
4	14
5	25

Table 2: Distribution of difficulty of questions as determined by the CollegeBoard.

One other dataset used to train the methods is human evaluation on how well a given answer choice fits in a sentence. To do this, fifty random questions of the 108 were selected. For each of these questions, five different sentences were generated by filling the blank(s) with one of the five answer choices such that each sentence represented a different answer choice. For example, for the first question in figure 1, we would generate the following sentences:

- The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including [outmoded] ones.
- The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including [figurative] ones.
- The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including [experimental] ones.
- The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including [cursory] ones.
- The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including [permanent] ones.

We then used Amazon's Mechanical Turk service,[4] an online platform for recruiting subjects to perform tasks, to collect the data. We set up a series of tasks asking people, or "turkers," how well they thought the word(s) inserted into the blank(s) belonged in a given sentence. Since this information is more nuanced than just which answer choice fit best, we hoped to use this data to develop more mature methods through which to answer the questions and also allow one to determine not only which answer choice is the most likely, but also which answer choices are the second, third, fourth, and fifth most likely to be chosen. To ensure quality of results, we used "master turkers," or people who have consistently provided high-quality responses, and had twenty

[4]https://www.mturk.com/mturk/welcome

turkers provide their evaluation for each sentence. An example of one of a questions a turker would be asked to answer displayed in Figure 3:

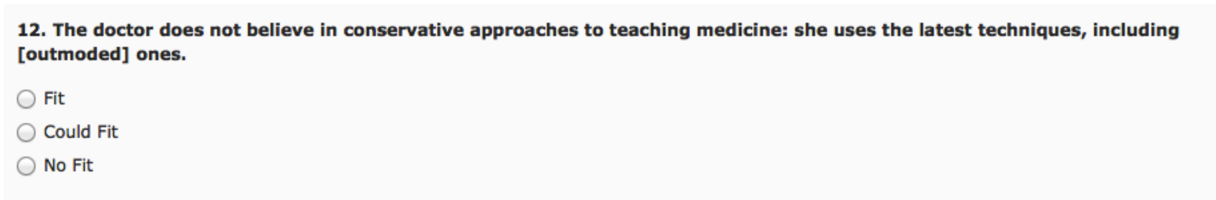

Figure 3: Sample of a question a turker would be asked to answer.

2.1. Data Preprocessing

Before the methods were trained and evaluated, we performed two preprocessing steps on both the corpus and the questions themselves.

The first preprocessing step was to remove commonly-used words, or stopwords, such as *the*, *is*, and *not*. It is true that stopwords can contain semantic meaning integral to answering a question. However, we decided to remove stopwords since all the methods used are "bag-of-words" models that do not take the order of the words into account, and stopwords do not contribute much information to the sentence meaning. Furthermore, although words such as *a* and *an* give key hints as to the qualities of the word following it (e.g. starting with a consonant or a vowel), the SAT sentence completion questions are crafted in a way such that these syntactic rules cannot be used to help determine the answer (see figure 4). The list of stopwords to be removed were obtained from the Python NLTK[5] library.

The second preprocessing step was to lemmatize the words. Lemmatization is the act of reducing inflected forms of a word to their base form. For example, the words *eat*, *ate*, and *eats*, are all inflected forms of the verb *eat*. If we were to understand how *eat* is used, it would be much more informative to consider the usages of *eat*, *ate*, and *eats* together rather than to only consider the usage of the specific form *eat*. Lemmatizing reduces the different forms of nouns, verbs, and adjectives to their base forms (e.g. *eat*, *ate*, and *eats* to *eat*), thus allowing more accurate analysis of

[5] http://www.nltk.org/

how they are used. To lemmatize the words, we used the WordNet Morphy lemmatizer contained in NLTK.

7. At bedtime the security blanket served the child as ------- with seemingly magical powers to ward off frightening phantasms.

(A) an arsenal (B) an incentive (C) a talisman (D) a trademark (E) a harbinger

Figure 4: Some questions will adjust the answer choices such that no syntactic clues can be used.

2.2. MSR

In 2011 Microsoft Research (MSR) released a set of sentence completion questions meant as a standard point of comparison for sentence completion question answering systems [24]. This data set consists of 1,040 sentence completion questions, each of which has five possible answer choices. Of the five answer choices, one is correct answer. The sentence completion questions are constructed based on sentences selected from five of Sir Arthur Conan Doyle's *Sherlock Holmes* novels. With the sentence completion questions, MSR also provides a corpus of 19th century novels off of which the methods can be trained. In addition to assessing our solvers on the SAT sentence completion questions, we assessed our final solvers on the MSR dataset as a basis of comparison.

3. Methodology

In the following section we discuss the intuition behind as well as the implementation of the five corpus-based methods explored in this study. Two of the methods, NPMI and co-occurrence frequencies, select their answer by observing the context surrounding the blank. The other three methods, LSA, CSKIP, and CBOW, select their answer by calculating the similarity between each answer choice and the words in the sentence.

3.1. Pointwise Mutual Information

First introduced by Church and Hanks in 1990 [3], Pointwise Mutual Information (PMI) is a measure of association between two events. More specifically, PMI measures how much the actual co-occurrence of two outcomes x and y differs from what we would expect if we assumed that the two outcomes were independent. PMI is expressed by the formula below:

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Positive values indicate that x and y co-occur more than expected while negative values indicate that when one event occurs, the other tends not to occur. As defined, PMI has an unbounded range, so we normalize the PMI using the following formula introduced by Bouma [1]:

$$NPMI(x,y) = -\left(\ln \frac{P(x,y)}{P(x)P(y)}\right) \Big/ \ln P(x,y) \tag{2}$$

This normalized PMI (NPMI) value guarantees that the values will be between -1 and 1 where NPMI equals 1 if the two terms only occur together, -1 if they only occur separately, and 0 if their occurrences are as expected under independence.

PMI/NPMI is used often to measure the similarity and co-occurrence between two words [1, 3, 18, 19]. In this study, NPMI is used to measure the co-occurrence frequency between each answer choice and the word that would be after it in the sentence.[6] If no word were after the blank in a sentence, then the word before the blank would be used. For each answer choice a score is thus calculated as follows:

$$Score(ans_i) = \begin{cases} NPMI(ans_i[0], w_1) & \text{if the question has one blank} \\ NPMI(ans_i[0], w_1) + NPMI(ans_i[1], w_2) & \text{if the question has two blanks} \end{cases} \tag{3}$$

[6]For example, in the first example in Figure 2, the NPMI would be found between each answer choice and *ones*.

where ans_i is the ith answer choice with $ans_i[0]$ being the word to be placed in the first blank and $ans_i[1]$ being the word to be placed in the second blank (if applicable). w_1 is the word after the first blank, and w_2 is the word after the second blank (if applicable). If there are no words after the second blank w_2, the word before the second blank is used instead. The probabilities are approximated based on relative word frequency in GloWbE, and Laplace smoothing is used to smooth the probabilities. The choice with the highest score is selected as the answer for each question.

3.2. Co-occurrence Frequencies

One common method that has been used in many word similarity problems is to look at word co-occurrence counts [2]. The idea behind this method is that the context of a word can be indicative of the word itself. However, the definition of what a "context" is can often vary. In this study, the "context" of a word is defined to be the five words to the right and the five words to the left of the word. Let $n(w, w_i)$ be the number of times the word w_i occurs in the context of w in the GloWbE corpus. For each word w a function f_w is created mapping each word in the vocabulary V to its relative co-occurrence frequency with w as follows:

$$f_w(w_i) = \frac{n(w, w_i)}{\sum_{w_j \in V} n(w, w_j)} \tag{4}$$

For each answer choice ans_i, a score is calculated as follows:

$$Score(ans_i) = \begin{cases} \sum_{w \in S} f_{ans_i[0]}(w) & \text{if the question has one blank} \\ \sum_{w \in S} \left(f_{ans_i[0]}(w) + f_{ans_i[1]}(w) \right) & \text{if the question has two blanks} \end{cases} \tag{5}$$

where S is the set of words in a sentence. The answer choice with the highest score is selected.

3.3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a vector-space model mapping different words to the concepts they are related to. LSA has already been shown to be effective in various question-answering applications including TOEFL exam questions [11] and biology multiple-choice questions [13]. In their 2012 paper, Zweig et al. found that LSA was the best individual method to answer SAT sentence completion questions, with an accuracy of 46% [25]. For this reason, we again consider the LSA metric here.

To perform LSA, we begin with a term-document matrix D representing how often each term appears in a document. The set of documents are obtained from the GloWbE. The term-document matrix is then reweighted using the TF-IDF weighting below as defined by Platt et al. [16] and used by Zweig et al. in their implementation of LSA [25]:

$$D_{ij} = \ln(f_{ij}+1)\ln(n/d_i) \tag{6}$$

where D_{ij} is the element in D corresponding to word i and document j, f_{ij} is the number of times word i appears in document j, n is the total number of documents, and d_i is the number of documents that contain word i. This reweighting accounts for the fact that terms that appear more often in a document are probably more important in determining the concepts in the document, but terms that appear in many documents (such as *the*) are perhaps not as defining.

After reweighting, singular value decomposition (SVD) is then performed on the matrix, resulting in three matrices U, S, and V. The resulting matrices from SVD are then truncated to a dimension d such that d is significantly smaller than the number of documents and the number of terms. The product of the three matrices $U_d S_d V_d^T$ represent an approximation to the original term-document matrix. In this case, a dimensionality of $d = 300$ was used, which was shown to work well on TOEFL exam questions in [11].

The most important aspect of LSA in this application is that the rows of $U_d S_d$, each of which represents a vector space for a given word, can be used to determine the similarity between two words. In this application, we determine the similarity between two words by taking the cosine similarity between the row vectors corresponding to the two words in $U_d S_d$. The cosine similarity between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined below:

$$\text{cossim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||||\mathbf{y}||} \tag{7}$$

A score for each answer choice is then determined by calculating the total word similarity as defined in [25]. Letting $LSAsim(w_0, w_1)$ be a function returning the cosine similarity between the row vectors corresponding to w_0 and w_1 in $U_d S_d$, we have:

$$Score(ans_i) = \begin{cases} \sum_{w \in S} LSAsim(ans_i[0], w) & \text{if the question has one blank} \\ \sum_{w \in S} (LSAsim(ans_i[0], w) + LSAsim(ans_i[1], w)) & \text{if the question has two blanks} \end{cases} \tag{8}$$

After computing each of these scores, the answer choice with the highest score is selected.

3.4. Word2Vec

Word2vec[7] is an implementation of two algorithms for finding vector representations of words. Created by Mikolov et al., it uses neural net language models to create word vectors based on the words surrounding a given word [14, 23]. The two language models in word2vec are the Continuous Bag-of-Words Model (CBOW) and the Continuous Skip-gram Model (CSKIP). Having already shown state-of-the-art performance in many applications [14], the algorithms in word2vec have shown much promise in various language-related applications. We apply word2vec to the SAT sentence completion questions in the following manner.

3.4.1. CBOW CBOW is a bag-of-words model trained to classify a word given the k words before and after a given word. Using a context window of 5, we train a CBOW model using word2vec on

[7] https://code.google.com/p/word2vec/

the GloWbE corpus. Letting $CBOWsim(w_0, w_1)$ be the cosine similarity between two word vectors in the CBOW model, a score for each answer choice is calculated as follows:

$$Score(ans_i) = \begin{cases} \sum_{w \in S} CBOWsim(ans_i[0], w) & \text{if the question has one blank} \\ \sum_{w \in S} (CBOWsim(ans_i[0], w) + CBOWsim(ans_i[1], w)) & \text{if the question has two blanks} \end{cases} \tag{9}$$

The answer choice with the highest score is selected.

3.4.2. CSKIP Instead of trying to predict a word given its context, CSKIP is a model trained by predicting the k words before and after a given word. Again using a window of 5, we train a CSKIP model using word2vec on the GloWbE corpus. Letting $CSKIPsim(w_0, w_1)$ be the cosine similarity between two word vectors in the CSKIP model, a score for each answer choice is calculated as follows:

$$Score(ans_i) = \begin{cases} \sum_{w \in S} CSKIPsim(ans_i[0], w) & \text{if the question has one blank} \\ \sum_{w \in S} (CSKIPsim(ans_i[0], w) + CSKIPsim(ans_i[1], w)) & \text{if the question has two blanks} \end{cases} \tag{10}$$

The answer choice with the highest score is selected.

3.5. Combination Methods

As shown by [20] and [25], combining different methods can lead to remarkably better results with respect to answering multiple choice questions. To combine the various methods, we performed a simple linear regression on the answer choice scores given by the various methods as well as different heuristics. Lasso linear regression was also performed since it generally returns a sparse coefficient vector (most of the coefficients are 0). For Lasso regression, we tuned the regularizer using 5-fold cross validation.

Since the AMT data provided us with human evaluation on how well a word fit in a sentence, for each answer choice we regressed its AMT score to a linear combination of the method scores and other heuristics. In addition to the scores predicted by NPMI, co-occurrence frequencies, LSA, CBOW, and CSKIP, the following predictors were also used.

3.5.1. 3-input LSA Used by Zweig et al. in [25], the score for each answer choice in 3-input LSA is a linear combination of the following three metrics:

- Total word similarity (see equation (8))
- Cosine similarity between the sum of the answer choice vectors and sum of word vectors in the question sentence
- Number of out-of-vocabulary terms in the answer

3.5.2. Other Heuristics In addition to those mentioned in 3-input LSA, we also tested the following heuristics:

- Sentence length
- Number of blanks
- Part(s) of speech before the blank(s)
- Distribution of parts of speech

4. Handling Low Word Frequencies: WordNet

One problem encountered when attempting to solve the SAT sentence completion questions is that some of the words being tested are used very infrequently in the English language. This phenomenon is reflected in the word counts of the different answer choices in the GloWbE corpus (see Figure 2). For example, the word *reviler* only appears twice in the GloWbE corpus. In an attempt to address this issue, we used the concept of synsets built into WordNet to expand the vocabulary considered.

WordNet is a lexical database that groups words into synonym sets called "synsets" [15]. Each synset contains words representing different ideas. For example, the word *board* may belong to two synsets—{*board*, *plank*} and {*board*, *committee*}—that represent the two different ways *board* can be used. To alleviate the issue of low word frequencies, instead of considering the similarities between each answer choice and the sentence, we consider the similarity between the synonyms in the synsets each answer choice belongs to and the sentence. For example, let us say we were analyzing the answer choice *placidity*, which only appears ten times in the GloWbE corpus. Instead of considering just the similarity between *placidity* and the sentence, we would consider the similarity between the synonyms of *placidity* as indicated by its synsets—{*placidity*, *quiet*, *placidness*, *tranquility*, ...}—and the sentence. Letting $\text{syn}(w)$ be the synonyms of w as defined above, we thus modify each of the aforementioned methods as follows:

4.1. NPMI

The bulk of this method remains the same, except when calculating NPMI and PMI in equations (1) and (2), instead of using the probabilities $P(x,y)$ and $P(x)$ where x is one of the answer choices, we use

$$P^*(x,y) = \sum_{w \in \text{syn}(x)} P(w,y) \tag{11}$$

and

$$P^*(x) = \sum_{w \in \text{syn}(x)} P(w) \tag{12}$$

respectively.

4.2. Co-occurrence Frequencies

Instead of using $f_w(w_i)$ in equation (5), $f_w^*(w_i)$ is used instead:

$$f_w^*(w_i) = \frac{\sum_{w^* \in \text{syn}(w)} n(w^*, w_i)}{\sum_{w^* \in \text{syn}(w)} \sum_{w_j \in V} n(w^*, w_j)} \tag{13}$$

4.3. LSA, CBOW, and CSKIP

In all three methods, to calculate the score of an answer choice, we find the cosine similarity between the vectors corresponding to the answer choice a and a word in the sentence w. Using WordNet, we modify this by finding the cosine similarity between $\text{syn}(a)$ and w. This is done by taking the mean of the vectors corresponding to the words in $\text{syn}(a)$ and finding the cosine similarity between the computed mean vector and the vector corresponding to w.

5. Experimental Results

5.1. Language Model Results

Table 3 summarizes the results obtained from using the NPMI, co-occurrence frequencies, LSA, CBOW, and CSKIP language models. In addition to showing the number of correct answers, the performance of the models were also analyzed in terms of the proportion of correctness across questions with one and two blanks and across varying difficulties of questions. Furthermore, for the questions, we also measured three other statistics. For each question that was answered incorrectly, we measured the error margin and the error rank. The error margin is a measure of how close the method was to selecting the correct answer by comparing the score of the correct answer and the answer choice selected by the method. The error rank similarly measures how close the method was to selecting the correct answer but instead does this by measuring the rank of the correct answer with respect to the other answers (did it have the second, third, fourth, or fifth best score). For questions that were answered correctly, we measured the correctness margin. This metric measures how by how much of a margin the method selected the correct answer by comparing the score of the correct answer and the answer choice with the second highest score. The formulas for the error margin and correctness margin for a question q are defined as follows:

$$\text{error margin} = \frac{qscore_1 - qscore[selectedans]}{qscore_1 - qscore_5} \tag{14}$$

$$\text{corectness margin} = \frac{qscore_1 - qscore_2}{qscore_1 - qscore_5} \tag{15}$$

where $qscore_i$ is the ith highest score an answer choice received for the question, and $qscore[selectedans]$ is the score given to the answer that the method selected for the question. Note that in the equation for correctness margin $qscore_1 = qscore[selectedans]$ since the method selected the correct answer, meaning that the correct answer choice had the highest score for the question. Table 3 displays the average error margin, error rank, and correctness margin for each method.

Method	**% Correct**	**% Incorrect by No. Blanks**		**% Incorrect by Difficulty**					**Avg. Error Margin**	**Avg. Error Rank**	**Avg. Correctness Margin**
		1	2	1	2	3	4	5			
NPMI	**30**	77	67	76	61	83	57	72	53	3.15	28
Co-occ. Freq.	**52**	50	46	43	56	53	43	44	67	3.17	44
LSA	**39**	63	58	48	67	73	64	52	54	3.17	35
CBOW	**48**	47	58	48	33	73	50	44	55	3.09	36
CSKIP	**48**	45	60	48	44	57	57	52	45	2.93	26

Table 3: Raw results for each of the methods.

Of these different methods, the co-occurrence frequencies had the best accuracy with 52% of the questions correct while NPMI had the worst accuracy with 30% correctness. For comparison, chance performance is 20%. Additionally, the co-occurrence frequency method had the best average correctness margin while CSKIP had the lowest average error margin.

In these results, one interesting phenomenon to note is that the performance of these methods seem uncorrelated with the number of blanks and the difficulty of each of the questions.

5.2. WordNet Expansion Results

Table 4 summarizes the results obtained from expanding the set of words considered using WordNet for each of the five measures. Table 5 shows a similar set of results, except when the WordNet expansion step was only applied to words that appeared fewer than 100 times in the corpus.

As one can see, using WordNet expansion either did not improve or actually substantially reduced the performance of all the methods. Applying WordNet expansion only to words that appear less frequently resulted in similar performance to that without WordNet expansion as in table 3. Due to this phenomena, we decided not to further explore using WordNet in the experiments below.

Method	% Correct	% Incorrect by No. Blanks		% Incorrect by Difficulty					Avg. Error Margin	Avg. Error Rank	Avg. Correctness Margin
		1	2	1	2	3	4	5			
NPMI	**27**	67	81	81	67	80	71	64	59	3.44	30
Co-occ. Freq.	**40**	64	72	57	61	50	64	72	53	3.02	52
LSA	**36**	62	67	57	61	70	86	52	57	3.10	40
CBOW	**46**	47	63	57	50	67	57	36	43	2.91	31
CSKIP	**44**	50	63	57	61	67	50	40	45	3.12	31

Table 4: Results from using WordNet expansion on all answer choices.

Method	% Correct	% Incorrect by No. Blanks		% Incorrect by Difficulty					Avg. Error Margin	Avg. Error Rank	Avg. Correctness Margin
		1	2	1	2	3	4	5			
NPMI	**29**	75	67	81	67	73	50	76	56	3.31	33
Co-occ. Freq.	**52**	52	44	43	50	47	50	52	66	3.13	45
LSA	**42**	60	56	48	61	63	64	56	58	3.22	36
CBOW	**45**	50	60	52	39	60	71	52	46	2.97	35
CSKIP	**52**	45	52	38	50	57	57	40	44	3.06	28

Table 5: Results from using WordNet expansion on answer choices with a frequency fewer than 100 in the GloWbE corpus.

5.3. Amazon Mechanical Turk Results

After collecting the turker responses on each of the questions, for each answer choice we had twenty evaluations as to whether the answer choice fit, could fit, or did not fit in the sentence. To quantify these measures, we assigned a score of 1 to every "fit" response, 0.5 to every "could fit" response, and 0 to every "no fit" response. We then averaged the twenty individual evaluations to determine an average score for each answer choice. The highest possible score for an answer choice is thus 1, indicating that people believed that the answer choice definitely fit in the sentence, and the lowest possible score is 0, indicating that people believed that the answer choice definitely could not fit in the sentence. An example of the scores given to the first question in figure 1 is shown in table 6.

As one can see, the human turkers performed very well on this specific question, with all the turkers believing that *experimental*, which was also the correct answer, fit well in the sentence. However, this data also provides us with valuable insight as to how well humans feel the other answer choices belong in the sentence, thus allowing us to discern that *figurative* is the second-best

Answer Choice	Average Score	Standard Deviation
outmoded	0.15	0.36
figurative	0.45	0.44
experimental	**1.00**	**0.00**
cursory	0.38	0.35
permanent	0.38	0.38

Table 6: Human evaluation of answers to the sentence: "The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including ——- ones." The correct answer ("experimental") is bolded.

option, with *cursory* and *permanent* tying for third, and *outmoded* being the worst option. It makes sense that the term *outmoded* would be the least likely to fit since it is antonymic to the idea in the question that "The doctor does not believe in conservative approaches."

Overall, the human turkers performed very well on the tasks. The correct answer had the highest average score for 92% (46/50) of the questions. Furthermore, of the 46 questions, the average score given to the correct answer is 0.90, and the average difference between the scores of the correct answer and the answer with the second-highest score is 0.35.

Table 7 shows the characteristics of the four other questions.

Question Difficulty	No. of Blanks	Correct Answer	Average Scores				
			A	B	C	D	E
3	1	A	**0.40**	**0.53**	0.40	0.38	0.23
4	1	C	**0.53**	0.35	**0.47**	0.28	0.23
5	1	A	**0.80**	0.68	0.50	0.60	**0.93**
5	1	D	**0.53**	0.40	0.30	**0.38**	0.35

Table 7: Characteristics of the questions for which the answer choice with the highest average score was not the correct answer.

The corresponding text of the four questions answered incorrectly is shown below (in the same order as the table):

```
The entrepreneur had a well-deserved reputation for -----,
having accurately anticipated many changes unforeseen by established
business leaders.
(A) prescience      (B) sincerity      (C) avarice
(D) complicity      (E) mendacity

Unlike sedentary people, ----- often feel a sense of rootlessness
instigated by the very traveling that defines them.
(A)  athletes       (B)  lobbyists       (C)  itinerants
(D)  dilettantes    (E)  idealists

The geologist speculated that eons ago, before the area was -----,
the present-day island was actually a hilltop in a vast forest.
(A) inundated      (B) situated      (C) rejuvenated
(D) supplanted     (E) excavated

Oren missed the play's overarching significance, focusing instead on
details so minor that they would best be described as -----.
(A)  pragmatic       (B)  indelible       (C)  moribund
(D)  picayune        (E)  impervious
```

As evidenced by the four questions highlighted in table 7, the turkers are not perfect in their evaluation in how well a word fits in the sentence. However, since different people have differing opinions on how well a question fits in a sentence, and due to the strong results of the turkers on the other questions, we felt that the data collected from Amazon Mechanical Turk well-represented on average how well a word would be perceived to fit in a sentence.

To assess how well each of the individual methods perform in terms of predicting how well each answer choice fits in a sentence, we found the Pearson correlation coefficient between the score given to each answer choice by the method and the corresponding human score. The Pearson correlation coefficient measures the linear dependence between two variables. The value of the coefficient is between 1 and -1 inclusive, where a coefficient of 1 indicates absolute positive correlation, 0 indicates no correlation, and -1 indicates absolute negative correlation. The results are given in table 8.

Method	Pearson's Correlation Coefficient
NPMI	0.15
Co-occ. Freq.	0.26
LSA	0.24
CBOW	0.35
CSKIP	0.39

Table 8: Correlation coefficient between method scores and human scores.

The correlation coefficients and the performance of the methods seem closely tied. The CBOW, CSKIP, and the co-occurrence frequencies methods had the highest correlation coefficients and correspondingly had the three highest performances on the questions. NPMI had the lowest correlation coefficient and had the lowest performance on the questions, and similarly with the LSA method. These results further corroborate the relative quality of the different methods as well as the quality of the Amazon Mechanical Turk data.

5.4. Combination Results

Previous studies have shown that combining different methods have given promising results in solving multiple-choice questions [20, 25]. We thus explored whether combining the various methods explored above would provide better results. To do this, we ran least squares and lasso linear regressions using the method scores as the regressors and the AMT results as the dependent variable. We thus trained a linear model on the 50 questions used in AMT and then tested the model on the remaining 58 questions. The results of the various combination methods are shown in table 9.

Of these various combination methods, a lasso regression of NPMI, co-occurrence frequencies, LSA, CBOW, and CSKIP performed the best, answering 59% of the questions correctly. Similarly, a lasso regression of NPMI, co-occurrence frequencies, LSA, CBOW, CSKIP, as well as the various heuristics described in the methodology section also answered 59% of the questions correctly. This value is greater than the current reported state-of-the-art performance of 53% [25].

Method	Least Squares		Lasso	
	R^2	Test Accuracy (%)	R^2	Test Accuracy (%)
A	0.06	43	0.05	40
B	0.09	40	0.09	40
C	0.16	57	0.16	**59**
D	0.17	50	0.14	53
E	0.30	36	0.15	**59**

A: 3-input LSA
B: NPMI + Co-occ. Freq. + LSA
C: NPMI + Co-occ. Freq. + LSA + CBOW + CSKIP
D: NPMI + Co-occ. Freq. + 3-input LSA + CBOW + CSKIP
E: NPMI + Co-occ. Freq. + LSA + CBOW + CSKIP + heuristics

Table 9: Results of combination methods.

5.5. MSR results

The results of running our various methods on the MSR dataset are shown in table 10. All the methods were trained on the corpus of 19th century novels provided by Microsoft while the combination method is the same model presented in the previous section due to the lack of human evaluation data on the MSR dataset.

Method	**SAT % correct**	**MSR % correct**
NPMI	30	34
Co-occ. Freq.	52	38
LSA	39	28
CBOW	48	47
CSKIP	48	49
NPMI + Co-occ. Freq. + LSA + CBOW + CSKIP	59	48

Table 10: Results of running methods on the MSR sentence completion questions.

On the MSR dataset, the NPMI, CBOW, and CSKIP methods performed similarly to how they performed on the SAT dataset whereas the other methods had a decrease in performance of about 10%.

6. Discussion

To better understand how the different methods work, we first note several qualities about our methods and the SAT sentence completion questions themselves. Since the SAT sentence completion questions are designed to be answerable without any prior knowledge, all the required semantic information required to answer a question is embedded in the question itself. To see this, let us consider the first question in figure 1, "The doctor does not believe in conservative approaches to teaching medicine: she uses the latest techniques, including ——- ones." The phrases *does not believe in conservative approaches* and *uses the latest techniques* indicate that the word in the blank should be a word that has a similar meaning to *latest* but perhaps has a connotation of being even newer and cutting-edge. This characteristic of the SAT questions most likely explains the poor performance of the NPMI method. The NPMI method only looks at a word adjacent to the blank. However, the context indicating which word belongs in the blank is embedded into the entire sentence, often several words away from the blank itself. Thus the NPMI method most likely does not have enough context to determine the proper word that belongs in the sentence. In contrast, the co-occurrence frequencies method focuses on the entire context surrounding each blank. This perhaps also explains why it has the best performance among the individual methods.

Of the methods, CBOW and CSKIP have the strongest correlation to human evaluation on the questions. On the SAT question set, they performed nearly as well as co-occurrence frequencies. Furthermore, they had the best performance on the MSR question set, and they are fairly robust methods as well—their performance on the MSR question set is similar to that on the SAT question set. This suggests that CBOW and CSKIP are strong candidates for future studies when addressing similar questions.

In the lasso regression combining NPMI, co-occurrence frequencies, LSA, CBOW, and CSKIP, it is of note that only the co-occurrence frequencies, CBOW, and CSKIP scores had nonzero coefficients. This indicates that among the five scores, these three scores are strong predictors for SAT sentence completion questions. Furthermore, it is interesting to note that the co-occurrence frequencies method examines the context surrounding the blank while CBOW and CSKIP analyzes the similarity between the words in the question and each answer choice. The state-of-the-art outcome from the combination of the three scores imply that the context and word-similarity methods can be effectively combined to enhance results.

In the MSR results, it is interesting to note that the co-occurrence frequency, LSA, and the combination method performed 10% worse on the MSR sentence completion questions than on the SAT sentence completion questions. For LSA, one possible explanation for its diminished performance is the sensitivity of LSA to various parameters [13]. Since we did not tune any of the parameters of LSA in this study, it is possible that the differing nature of the MSR questions resulted in a diminished performance. For co-occurrence frequencies, one possible reason for its diminished performance is the fact that for the MSR questions, the context of the question does not necessarily provide enough information to answer the question. Take, for example, the following question:

```
Presently he emerged, looking even more _____ than before.
    a) instructive
    b) reassuring
    c) unprofitable
    d) flurried
    e) numerous
```

In this question, it is not absolutely clear without looking at the answer choices what the meaning of the word in the blank should be. Even with the answer choices, the correct answer could still be debated. Thus, without the contextual clues necessary to make its decision, the co-occurrence frequency method did not perform as well. For the combination method, it is most likely the degraded method of the co-occurrence frequency method and the fact that the combination method was trained on the SAT sentence completion questions that lead to diminished results.

7. Related Work

Currently, the state-of-the-art performance in answering SAT sentence completion questions is from the 2012 paper by Zweig et al. [25]. In this paper they explored various local and global information methods through which to answer the questions, including various n-gram models, a recurring neural net model, and various LSA models. They found an optimal correctness of 53% by using a linear combination of the outputs of their Good-Turing smoothed n-gram and LSA total similarity models. Beyond this, there has been no significant work done specifically on SAT sentence completion questions. However, there has been extensive work performed on TOEFL synonym questions and SAT analogy questions. These studies looked at many of the methods considered here such as NPMI, LSA, and various co-occurrence measures [2, 11, 18, 19]. Furthermore, studies have also been performed exploring the different ways of combining the output of various methods [20], although in this study we only considered the baseline linear combination.

Since 2011, there have also been several studies related to the MSR dataset. Of these, perhaps the most notable is a recent 2014 study showing a state-of-the-art performance of 87.4% on the dataset [12]. They achieved this performance by considering various n-gram smoothing techniques in conjunction with using the Google Web1T N-gram Count corpus, a database of ~1 trillion different English 5-grams based on text from the web.[8]

8. Conclusion and Future Work

Through this study, we assessed the ability of various context-based and similarity-based methods to answer SAT sentence completion questions. We found that the context-based co-occurrence frequencies method performed the best with 52% correctness, and that combining co-occurrence frequencies with the similarity-based CBOW and CSKIP methods resulted in a state-of-the-art 59% correctness. As a basis of comparison, we ran our methods on the MSR sentence completion questions. Most methods performed similarly, although the co-occurrence frequencies, LSA, and combination method did not, most likely due to the differing nature of the questions.

From this study, we feel that there are various avenues of further work to be performed. First, we note that each solver generally performs well on different questions. As shown in figure 5, most of the questions had between one and three solvers answer it correctly. There were only eight questions that all the methods answered correctly and sixteen that none answered correctly. The diversity in

[8] https://catalog.ldc.upenn.edu/LDC2006T13

the questions answered by each method seems to indicate that certain methods are better at solving certain questions than others. If the characteristics of the questions that each method answers well can be identified, these characteristics can be leveraged to potentially create a conglomerate method with even better performance.

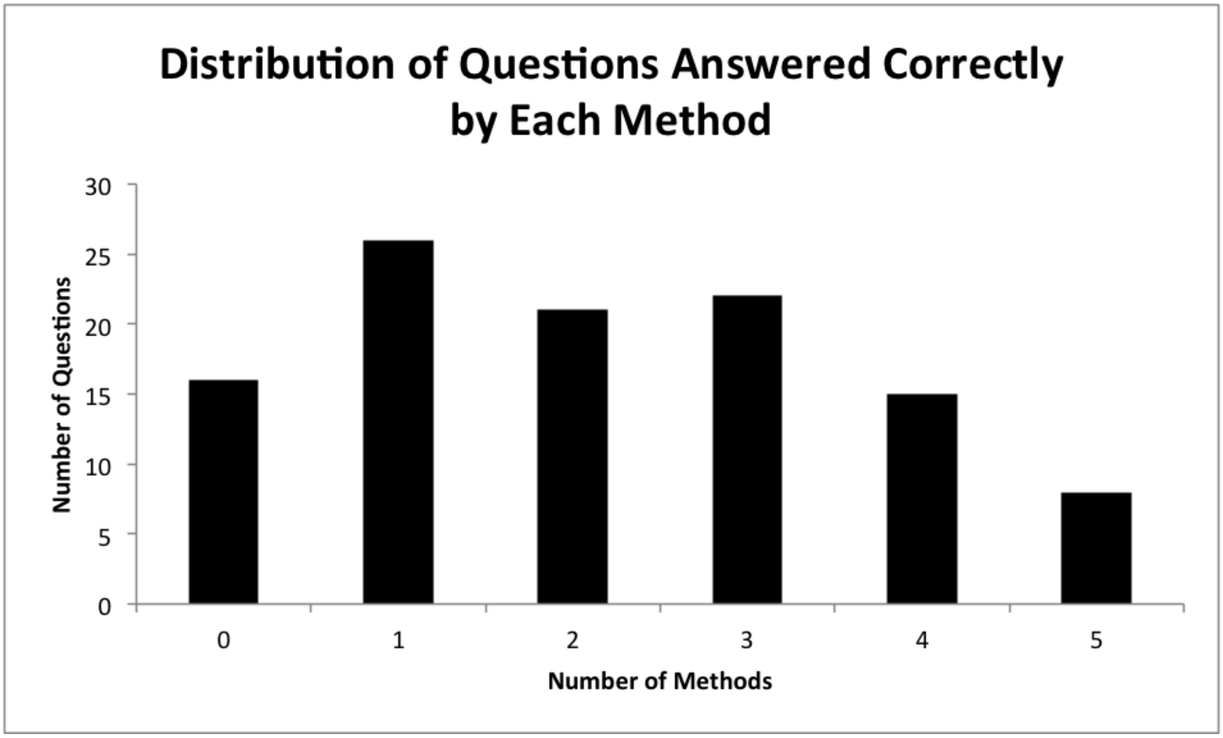

Figure 5: Distribution of SAT questions answered correctly.

The fact that none of the methods was able to answer 16 (15%) of the questions correctly also seems to indicate a limitation of context- and similarity-based methods. For example, one of the questions that all the methods answered incorrectly is:

```
The play closed after only a week because critics gave the
performance ----- reviews.
    (A)  innocuous
    (B)  caustic
    (C)  rave
    (D)  gaudy
    (E)  contrite
```

The correct answer to this question is *caustic*. However, all the methods selected as their top answer its antonym, *rave*. The reason for this is most likely due to the fact that antonyms often appear in similar contexts [17, 21], thus the context-based and similarity-based methods have a hard time distinguishing between antonyms even though the semantic meaning of antonyms are completely opposite of each other. Thus one other possible avenue of exploration is to study other methodologies through which to determine how well a word fits in a sentence, such as by observing the syntax tree of the sentence to help determine what type of word belongs in the blank.

One way we hope to further explore the methods studied here is parameter tuning. Currently, we fixed many of the parameters in our various methods, such as the context window in co-occurrence frequencies and the number of dimensions in LSA. However, it is possible that these parameters can significantly change the performance of the methods. For example, in [11], Landaeur and Dumais showed that there is a small range of dimensionality values for which LSA would have a sharp peak of performance on TOEFL questions. We feel that tuning these parameters could have a noticeable effect in the performance of the methods, but that the parameters used here are at the very least a representative benchmark of the relative performances of the different methods.

Acknowledgments

First, I would like to thank the Princeton School of Engineering and Applied Sciences without whose generous funding this project would not be possible. I would also especially like to thank Professor Christiane Fellbaum for her invaluable guidance and support throughout the research process.

References

[1] G. Bouma, “Normalized (pointwise) mutual information in collocation extraction,” *Proceedings of GSCL*, pp. 31–40, 2009.

[2] J. A. Bullinaria and J. P. Levy, “Extracting semantic representations from word co-occurrence statistics: A computational study,” *Behavior research methods*, vol. 39, no. 3, pp. 510–526, 2007.

[3] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[4] Educational-Testing-Service. (2003) Sat preparation booklet. Available: http://www.collegeboard.com/prod_downloads/sat/satguide/SAT_Full.pdf

[5] Educational-Testing-Service. (2004) Sat preparation booklet. Available: http://www.collegeboard.com/prod_downloads/sat/satguide/SAT_full_0405.pdf

[6] Educational-Testing-Service. (2005) Sat preparation booklet. Available: http://www.collegeboard.com/prod_downloads/prof/counselors/tests/sat/sat-prep-book-stu.pdf

[7] Educational-Testing-Service. (2007) Sat preparation booklet. Available: http://www.collegeboard.com/prod_downloads/prof/counselors/tests/sat/2007-08_sat_preparation_booklet.pdf

[8] Educational-Testing-Service. (2013) Sat preparation booklet. Available: https://satonlinecourse.collegeboard.org/SR/digital_assets/assessment/pdf/F4D31AB0-66B4-CE32-00F7-F5405701F413-F.pdf

[9] Educational-Testing-Service. (2014) Sat preparation booklet. Available: https://satonlinecourse.collegeboard.org/SR/digital_assets/assessment/pdf/0833A611-0A43-10C2-0148-CC8C0087FB06-F.pdf

[10] A. Jobbins *et al.*, "Postprocessing for ocr: Correcting errors using semantic relations," in *LEDAR. Language Engineering for Document Analysis and Recognition, AISB 1996 Workshop, Sussex, England*, 1996.

[11] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[12] K. Lee, G. G. Lee *et al.*, "Sentence completion task using web-scale data," in *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*. IEEE, 2014, pp. 173–176.

[13] A. Lifchitz, S. Jhean-Larose, and G. Denhière, "Effect of tuned parameters on an lsa multiple choice questions answering model," *Behavior research methods*, vol. 41, no. 4, pp. 1201–1209, 2009.

[14] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] G. A. Miller *et al.*, "Introduction to wordnet: An on-line lexical database*," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[16] J. C. Platt, K. Toutanova, and W.-t. Yih, "Translingual document representations from discriminative projections," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 251–261.

[17] T. Tanaka, "Measuring the similarity between compound nouns in different languages using non-parallel corpora," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 1–7.

[18] E. Terra and C. L. Clarke, "Frequency estimates for statistical word similarity measures," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 165–172.

[19] P. Turney, "Mining the web for synonyms: Pmi-ir versus lsa on toefl," 2001.

[20] P. Turney *et al.*, "Combining independent modules to solve multiple-choice synonym and analogy problems," 2003.

[21] X. Wei *et al.*, "Context sensitive synonym discovery for web search queries," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1585–1588.

[22] M. L. Wick, M. G. Ross, and E. G. Learned-Miller, "Context-sensitive error correction: Using topic models to improve ocr," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 1168–1172.

[23] Y. YUAN *et al.*, "A new study based on word2vec and cluster for document categorization," *Journal of Computational Information Systems*, vol. 10, no. 21, pp. 9301–9308, 2014.

[24] G. Zweig and C. J. Burges, "The microsoft research sentence completion challenge," Technical Report MSR-TR-2011-129, Microsoft, Tech. Rep., 2011.

[25] G. Zweig *et al.*, "Computational approaches to sentence completion," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 601–610.