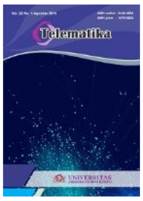

Terbit *online* pada laman web jurnal :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Terakreditasi Sinta "3" KEMENRISTEKDIKTI, No. 21/E/KPT/2018

UNIVERSITAS
AMIKOM PURWOKERTO

Synonym Measurement Through Semantic Similarity Using the SOC-PMI Method

Uswatun Hasanah[1], Bambang Pilu Hartato[2], Mitra Yulianti[3], Saeful Haq Faruqi[4]

[1,2,4]Department of Information Technology
[3]Department of Informatics
Faculty of Computer Science
Universitas Amikom Purwokerto
E-mail: uswatun_hasanah@amikompurwokerto.ac.id[1], bambang.pilu@amikompurwokerto.ac.id[2], mitrayulianti98@gmail.com[3], ruqisaeful@gmail.com[4]

ARTICLE INFO

History of the article:
Receive December 16, 2019
Revised January 16, 2020
Received February 24, 2020
Online February 28, 2020

Keywords:
SOC-PMI
Semantic Similarity
Synonym
Corpus-based Method

Kata Kunci:
SOC-PMI,
Kemiripan Semantik,
Sinonim,
Metode Berbasis Korpus

Correspondence:
Telephone: +62 (857) 12190708
E-mail:
uswatun_hasanah@amikompurwokerto.ac.id

ABSTRACT

Measurement of synonyms can be an essential task in measuring word similarity. This work cannot be done syntactically but must dig deeper into its semantics. Semantic relations can be anything, such as synonyms, antonyms, hyponymy, homonymy, and polysemy. This research works on finding synonym values using the Second Order Co-occurrence Pointwise Mutual Information (SOC-PMI) method. The data used are 30 questions on the TOEFL exam. Each question consists of one word as a question and four reference answers as alternative answers. The results show very low accuracy (30%) since there are only 9 out of 30 answers that show the synonym. Besides, the LCS method was also tested to get a character-based similarity score. LCS method can achieve a higher similarity score of 43.33%. Finally, the idea of the hybrid method by combining character-based and semantic-based methods can be considered in longer words to produce a fairer similarity score.

ABSTRAK

Pengukuran sinonim dapat menjadi pekerjaan yang penting dalam mengukur kemiripan kata. Pekerjaan ini tidak dapat dilakukan secara sintaksis, tetapi harus dilakukan dengan menggali lebih dalam tentang semantiknya. Hubungan semantik dapat berupa apa saja, seperti sinonim, antonim, hiponim, homonim, dan polisemi. Penelitian ini berusaha untuk menemukan nilai-nilai sinonim menggunakan metode Second Order Co-occurrence Pointwise Mutual Information (SOC-PMI). Data yang digunakan adalah 30 pertanyaan pada ujian TOEFL. Setiap pertanyaan terdiri dari satu kata sebagai pertanyaan dan empat jawaban referensi sebagai jawaban alternatif. Hasil menunjukkan nilai akurasi yang sangat rendah (30%) karena hanya ada 9 dari 30 jawaban yang benar-benar menunjukkan sinonim. Selain itu, metode LCS juga diuji untuk mendapatkan skor kemiripan berdasarkan karakternya. Metode LCS mampu mencapai skor kemiripan yang lebih tinggi yaitu 43,33%. Akhirnya, gagasan metode hybrid dengan menggabungkan metode berbasis karakter dan metode berbasis semantik semantik dapat dipertimbangkan untuk kata-kata yang lebih panjang agar menghasilkan skor kesamaan yang lebih adil.

INTRODUCTION

Two concepts or words can be related (or not) by expressing their semantic relatedness (Islam & Inkpen, 2006). The relation of meaning is described by the semantic relationship between one entity and another. In linguistics, entities can be words, phrases, clauses, or sentences. The relationship of meaning to

http://dx.doi.org/10.35671/telematika.v13i1.941

linguistics includes the similarity of meanings (synonyms), contradictory meanings (antonyms), coverage of meaning (hyponymy), doubling of meaning (homonymy), or excess of meaning (polysemy) (Parera, 2004). Synonyms refer to terms that can be used to describe a particular entity; for example, the entity "Holland" can refer to "Netherlands." In natural language processing applications, entity synonyms play an essential role. Some examples of its application are text summarization (Alguliyev, Aliguliyev, Isazade, Abdi, & Idris, 2017; Barzilay & Elhadad, 1999), query expansion (Aronson & Rindflesch, 1997; Díaz-Galiano, Martín-Valdivia, & Ureña-López, 2009), reformulation (Plovnick & Zeng, 2004), paraphrase detection, and question answering (Ferrucci, 2012).

This study aims to measure the similarity of synonyms by knowing the value of semantic similarity. Since semantic similarity can be of various types, this research limits only the synonym similarity. In research conducted by (Ullmann, 1964) in (Djajasudarma, 1993), synonyms are divided into nine types as follows: 1) Synonyms in which one of its members has a more general meaning, 2) Synonyms in which one of its members has more intensive elements of meaning, 3) Synonyms where one of the members emphasizes emotive meaning, 4) Synonyms where one of the members is reproachful or not justifying, 5) Synonym where one of the members becomes a field term specific, 6) Synonyms where one of its members is more widely used in a variety of written languages, 7) Synonyms where one of its members is more commonly used in conversational languages, 8) Synonyms where one of the members is used in childhood language, 9) Synonyms where one of the members is usually used in certain areas.

The method used in this study considers semantic similarity in measuring synonym similarity. The method, namely, Second-Order Co-occurrence Pointwise Mutual Information (SOC-PMI), was conceived by (Islam & Inkpen, 2006) and has been used in a variety of natural language processing applications. Even though the method focuses on semantic similarity, this research focuses on synonyms, which are one of a series of semantic elements.

RESEARCH METHODS

1. Semantic Similarity

Humans, with their common sense, can recognize the interrelation of a pair of words in various ways. For humans, it is not difficult to judge the relationship between apples and oranges, rather than apples and toothbrushes (Islam & Inkpen, 2006). Semantics can be used in two mechanisms, namely in the detection of similarities and differences (Frawley, 2013). During this time, applications in natural language processing have used semantic similarity measurements, such as in the construction of automated thesaurus (Grefenstette, 1993)(D. Lin, 1998)(Li, Abe, World, & Partnership, 1998), automatic indexing, text annotations and document summarizing (C. Lin, Hovy, & Rey, 2003), text classification, word sense disambiguation (Li et al., 1998)(Lesk, 1986)(Yarowsky, 1992), information extraction and information retrieval (Buckley, Salton, Allan, & Singhal, 1995)(Vechtomova & Robertson, 2014)(Xu & Croft, 2000).

2. SOC-PMI

The Second Order Co-occurrence Pointwise Mutual Information method, or from now on referred to as the SOC-PMI method, is a method developed from the predecessor algorithm called PMI-IR. PMI-IR is proposed by (Turney, 2001), and uses the AltaVista Advanced Search query syntax to calculate probabilities. PMI-IR is a simple method intended to recognize synonyms, using Pointwise Mutual Information as written in Equation (1) below:

$$\mathrm{score}(choice_i) = \mathrm{p}(problem\ \&\ choice_i)\ /\ \mathrm{p}(choice_i) \quad (1)$$

where, $\{choice_1, choice_2, \ldots, choice_n\}$ represent the alternatives from problem word $problem$, while probability that problem and $choice_i$ co-occur stated with $\mathrm{p}(problem\ \&\ choice_i)$. Another variation of this equation is based on the closeness of the pair in the document, considering antonyms, and considering the context.

Through the principle of probability PMI-IR, (Islam & Inkpen, 2006) formulate the Pointwise Mutual Information that can be shown by Equation (2) as follows:

$$f^{pmi}(t_i, W) = \log_2 \frac{f^b(t_i, W)\times m}{f^t(t_i) f^t(W)} \quad (2)$$

W is targeted word, while $f^t(t_i)$ is a type of frequency function, and $f^b(t_i, W)$ is a bigram frequency function. Then, the total number of tokens in corpus C represented by m. Furthermore, $\beta - PMI\ summation$ functions for W_1 and W_2 are defined in Equation (3) and (4):

$$f^\beta(W_1) = \sum_{i=1}^{\beta_1}(f^{pmi}(X_i, W_2))^\gamma \quad (3)$$

$$f^\beta(W_2) = \sum_{i=1}^{\beta_2}(f^{pmi}(Y_i, W_1))^\gamma \quad (4)$$

Finally, the PMI semantic similarity function between the two words W_1 and W_2 is shown by the following Equations (5):

$$Sim(W_1, W_2) = \frac{f^\beta(W_1)}{\beta_1} + \frac{f^\beta(W_2)}{\beta_2} \quad (5)$$

The value β is related to the number of times the word W appears in the corpus. The β value is defined in the following Equation (6):

$$\beta_1 = (\log\ (f^t(W_i)))^2 \frac{(\log_2(n))}{\delta}, \quad i = 1{,}2 \quad (6)$$

Where δ is constant and in research conducted by (Islam & Inkpen, 2006) the value δ = 6.5 is determined. The value δ depends on the size of the corpus. The smaller the corpus used, the smaller the value of δ.

3. Data

In this research, synonym similarity is obtained from two pairs of words with the semantic meaning approach. The data used is part of the TOEFL test which deals with synonyms. Data are collected from lessons 1-3 in the TOEFL exercise book written by (Matthiesen, 2017). A total of 30 multiple choice questions were used, and each question had four alternative answers. The following is an example of the data used:

Choose the synonym of *appealing*:

(A) refined

(B) encouraging

(C) alluring

(D) popular

The answer key provided by the book will provide information that the synonym of the word "appealing" is "alluring," in which the two words have the same meaning of the word "attractive." Besides, these words also have other synonyms, such as "interesting," "enticing," "catchy," and "catching."

http://dx.doi.org/10.35671/telematika.v13i1.941

4. Research flow

This research goes through the steps shown in Figure 1 below:

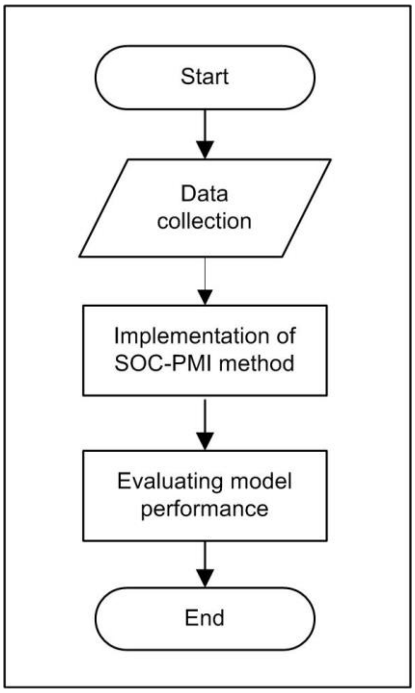
Start
Data collection
Implementation of SOC-PMI method
Evaluating model performance
End

Figure 1. Research steps

The first step is to collect data, as described in the previous section. The data does not experience any pre-processing techniques. In this case, the similarity data will be directly measured using the SOC-PMI method. The library obtained from https://github.com/pritishyuvraj/SOC-PMI-Short-Text-Similarity- is used to measure the semantic similarity. In the library, there are at least three algorithms included, where the three methods are Hybrid methods named Semantic Text Similarity (STS) (Islam & Inkpen, 2008). In the library, there are at least three algorithms included, where the three algorithms are Hybrid methods named Semantic Text Similarity (STS). However, in this study, only the SOC-PMI algorithm was taken and used. This method includes the NLTK library and also uses WordNet as the dictionary. Wordnet is an extensive semantic network in which there are words and groups of words that are connected lexically and conceptually, which are represented by arc labeled (Fellbaum, 2006).

Furthermore, after the SOC-PMI value of each possible answer is obtained, an evaluation of the method's performance is carried out by finding a match between the two answers, both the predicted answer and the actual answer. We also apply another method for comparison. We use a character-based method called Longest Common Subsequence. We use this method because it is not possible to implement string-based methods with questions in the form of word synonyms, since the words used are clearly different.

RESULTS AND DISCUSSION

The results and discussion of this paper can be seen as follows:

1. Results

The synonym question data in lessons 1, 2 and 3 collected and will measure the semantic similarity. Table 1 illustrates an example of the semantic similarity measurement results using the SOC-PMI method in question Number 5, Lesson 2:

Table 1. Word example using SOC-PMI method with synonym value

Question	*Answer*	*SOC-PMI Score*
appealing	refined	0.13333
	encouraging	0.14286
	alluring	**0.84211**
	popular	0.15385

Based on the results obtained in Table 1, it can be seen that the word "alluring" has the closest semantic relation to the word "appealing" with a similarity value of 0.84211. Furthermore, the word "alluring" in bold indicates that the word is the actual answer. In the end, the whole data is also measured, and the results obtained as shown in Table 2, Table 3, and Table 4.

Table 2. Results in lesson 1

No	*Question*	*Answer*	*SOC-PMI Score*	*LCS Score*
1	widely	**broadly**	0	**0,46154**
		abroad	0	0,16667
		secretly	0	0,42857
		truly	0	0,36364
2	autonomous	**independent**	0	0,09524
		sudden	0	0,25000
		international	0	0,34783
		abrupt	0	**0,37500**
3	advice	acclaim	**0,66667**	0,30769
		attention	0,30769	0,26667
		suggestion	0,30769	0,12500
		praise	**0,66667**	**0,50000**
4	attractive	**appealing**	0	0,31579
		adverse	0	0,35294
		arbitrary	0	**0,42105**
		perfect	0	0,35294
5	disapproval	attraction	0,13333	**0,28571**
		attention	0,28571	0,20000
		objection	0,28571	0,20000
		persistence	**0,375**	0,18182
6	haphazardly	suddenly	0	0,31579
		secretly	0	0,31579
		carelessly	0	**0,38095**
		constantly	0	0,28571
7	constant	disruption	**0,35294**	0,33333
		acceptable	0	0,33333
		abrupt	0	0,28571
		persistent	0	**0,44444**
8	perfect	attractive	0	**0,35294**
		ideal	**0,26667**	0,16667
		actual	0	0,30769
		abrupt	0	0,30769
9	unfavorably	attractively	0	0,43478
		haphazardly	0	0,36364
		acceptably	0	0,47619
		adversely	0	**0,50000**
10	disturbing	perfect	0,16667	0,11765
		disruptive	**0,4**	**0,50000**
		persistent	0,4	0,40000
		attractive	0,4	0,30000

Table 3. Results in lesson 2

No	*Question*	*Answer*	*SOC-PMI Score*	*LCS Score*
1	inspire	celebrate	**0,28571**	0,25000
		attract	0,22222	0,14286
		encourage	0,25	**0,37500**
		appeal	0,14286	0,30769
2	advantage	**benefit**	0,26667	0,25000
		persistence	0,4	0,20000
		nimbleness	0,30769	0,21053
		allure	**0,66667**	**0,26667**
3	fragile	modern	0	0,15385
		famous	0	0,30769
		allowable	0	0,37500
		frail	0	**0,83333**
4	contemporary	timing	0,15385	0,22222
		current	**0,28571**	**0,31579**
		well-known	0	0,18182
		perfect	0,14286	0,21053
5	appealing	refined	0,13333	0,37500
		encouraging	0,14286	0,50000
		alluring	**0,84211**	**0,58824**
		popular	0,15385	0,37500
6	renown	unknown	**0,14286**	**0,61538**
		celebrated	0	0,25000
		adverse	0	0,30769
		disapprove	0	0,25000
7	worthwhile	**rewarding**	0	0,31579
		acceptable	0	0,30000
		agile	0	**0,40000**
		permitted	0	0,31579
8	vigorous	attractive	0	0,11111
		beautiful	0	0,23529
		energetic	0	0,11765
		advantageous	0	**0,50000**
9	refine	persist	**0,25**	0,30769
		value	0,13333	0,18182
		perfect	0,15385	**0,46154**
		divide	0,16667	0,40000
10	distribute	disappoint	0,25	0,50000
		disrupt	0,22222	**0,70588**
		discourage	0,22222	0,50000
		dispense	**1**	0,44444

Table 4. Results in lesson 3

No	*Question*	*Answer*	*SOC-PMI Score*	*LCS Score*
1	indispensable	abrupt	0	0,21053
		abroad	0	0,21053
		vital	0	**0,33333**
		frail	0	0,22222
2	restore	appeal	0,14286	0,15385
		revitalize	**0,22222**	**0,47059**
		attract	**0,22222**	0,28571
		disrupt	0,2	0,28571
3	conform	annoy	**0,4**	0,33333
		divide	0,2	0,00000
		encourage	**0,4**	**0,37500**
		adapt	**0,4**	0,00000
4	notice	**observe**	0	0,30769
		refine	0	0,33333

http://dx.doi.org/10.35671/telematika.v13i1.941

		distribute	0	**0,37500**
		analyze	0	0,30769
5	current	energetic	0	**0,37500**
		ideal	**0,13333**	0,16667
		ongoing	0	0,14286
		intense	0	0,28571
6	observe	alter	**0,33333**	**0,33333**
		notice	0,16667	0,30769
		anticipate	0,25	0,11765
		modify	0,28571	0,15385
7	intense	**strong**	0	0,30769
		intolerant	0	**0,58824**
		vitally	0	0,28571
		allowable	0	0,12500
8	enrich	alter	**0,33333**	0,36364
		dispense	0,25	0,28571
		disrupt	0,22222	0,15385
		enhance	0,2	**0,46154**
9	unbearable	inspiring	0	0,21053
		unfavorable	0	**0,76190**
		intolerable	**1**	0,66667
		ancient	0	0,23529
10	proposal	question	0,28571	**0,12500**
		attention	**0,33333**	0,11765
		benefit	0,28571	0,00000
		suggestion	**0,33333**	0,11111

Based on the results shown in Table 2, Table 3, and Table 4, several things must be considered. First, some vocabularies do not show any semantic relations. Figure 2 illustrates the distribution of words that have semantic relations and those without semantic relations.

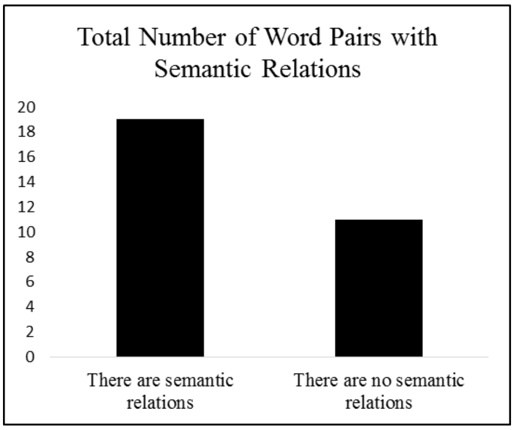

Figure 2. Total number of word pairs with semantic relations

In the end, the accuracy of the values generated by the SOC-PMI method with the actual answers is also measured. By reviewing the SOC-PMI values generated in Table 2, Table 3, and Table 4, it can be seen that there are 9 correct answers and 21 missed answers. Futhermore, the result for accuracy values are:

$$Accuracy = \frac{9}{30} \times 100\% = 30\%$$

Meanwhile, using the LCS method we obtain the following accuracy result:

$$Accuracy = \frac{13}{30} \times 100\% = 43.33\%$$

Unfortunately, it can be seen that the results are not satisfying results. The discussion section will explain the phenomena and analyze what factors influence the results and how this method can be used in the future.

2. Discussion

In this section, the results obtained are then analyzed. First, it should be noted that the SOC-PMI method is not evaluating semantic similarities based on synonymous rules. The SOC-PMI method considers the semantic relations between one pair of words, where semantic meaning can be anything. They can be synonymous, antonym, hyponymy, hypernimic, polysemic, or just connected to a certain hierarchy. Furthermore, this method takes into account how a pair of words meet in the same context. At this point, the frequency with which each word appears in the same context window greatly influences the results of semantic similarity. For example, the word pairs "computer" and "machine" will have more similar semantic relations (i.e. 0.94118) than "computer" and “keyboard" (0.82353), "computer" and "portable" (0.76190), and "computer" and "RAM" (0.70000).

High or low semantic similarity value is determined by the frequency of occurrence of the two words together in the context window. Even though the completeness of the word dictionary will also affect the results of semantic similarity. In the previous section, there were eleven questions for which the alternative answers did not have any semantic relations. This can happen for two reasons. First, the two words do not appear at all in the corpus, or it can only appear one of them without being followed by the next word. Secondly, the two words do exist in the corpus but do not appear in the same context. Therefore, the SOC-PMI similarity value cannot be obtained. In the case of *Netherlands* - *Holland*, *computer* - *keyboard*, *computer* - *machine*, or *mommy* - *daddy*, the SOC-PMI method might be able to provide competitive results, depending on the size of the corpus used. However, if the corpus is not able to represent words that are not commonly used, that will be another problem.

Here, the LCS method may have better performance. However, the LCS method ignores the semantic meaning of words because it considers the presence or absence of a character in the two words being compared. Sometimes the LCS method can give a higher score even though the characters are reversed. In this case, the idea of combining character-based and semantic-based methods can be considered in longer words, i.e., phrases or sentences. In the end, the hybrid method can be considered to produce a fairer similarity score. For example, we can give each word score weighting for the SOC-PMI and LCS values. The word "restore" has the synonym word "revitalize" where the SOC-PMI method gives a score of 0.22, and the LCS method gives a score of 0.47. If we give each method a weight of 0.5, we will get a final similarity score of 0.35.

For future NLP works involving word similarity factors, it can be concluded that the SOC-PMI method is not specifically recommended for personal use. As in this case, it is used to determine the synonymity of words. It would be wise to use the SOC-PMI method together with other methods (so that it will become a new hybrid method). This idea starts from the perspective that the relation of semantic meaning can be in any form. Thus, considering the possibility of syntactic similarity will be wiser and more objective on tasks involving more general similarities.

CONCLUSIONS AND FUTURE WORKS

Finally, a conclusion can be drawn from the results and discussion in the previous section. Firstly, the SOC-PMI method is not suitable for determining specific semantic meanings, such as the case of synonyms between words. Because semantic relations can be of any type, depending on the frequency with which the two words occur together in the context window in the corpus. Secondly, the SOC-PMI method might perform well when measuring the semantic similarity of commonly used words, but this does not necessarily apply to TOEFL synonym questions because they have vocabulary lists that are sometimes "not common." In the end, the SOC-PMI method might work better when combined with other methods. However, the ability of WordNet will be a new challenge for other types of languages.

ACKNOWLEDGEMENT

The authors would like to thank Universitas Amikom Purwokerto for funding this research.

REFERENCE

Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2017). A model for text summarization. *International Journal of Intelligent Information Technologies (IJIIT)*, *13*(1), 67–85.

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium* (p. 485).

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. *Advances in Automatic Text Summarization*, 111–121.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. *NIST Special Publication Sp*, 69.

Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, *39*(4), 396–403.

Djajasudarma, T. F. (1993). *Semantik I: Pengantar ke Arah Ilmu Makna*. *Eresco 145*. Bandung.

Fellbaum, C. (2006). WordNet(s). In *Encyclopedia of Language & Linguistics (Second Edition)* (pp. 665–670).

Ferrucci, D. A. (2012). Introduction to “This is Watson.” *IBM Journal of Research and Development*, *56*(3.4), 1.

Frawley, W. (2013). *Linguistic semantics*. Routledge.

Grefenstette, G. (1993). Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques. In *Ninth Annual Conference of the UW Centre for the New OED and Text Research*.

Islam, A., & Inkpen, D. (2006). Second order co-occurrence PMI for determining the semantic similarity of words. In *LREC* (pp. 1033–1038). https://doi.org/10.1145/1376815.1376819

Islam, A., & Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(2), 1–25. https://doi.org/10.1145/1376815.1376819

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24–26).

Li, H., Abe, N., World, R., & Partnership, C. (1998). Word clustering and disambiguation based on co-occurrence data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 749–755).

Lin, C., Hovy, E., & Rey, M. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 71–78).

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics* (pp. 768–774).

Matthiesen, S. J. (2017). *Essential Words for the TOEFL*. Simon and Schuster.

Parera, J. D. (2004). Teori Semantik [Semantic Theory]. *Jakarta: Erlangga*.

Plovnick, R. M., & Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of Medical Internet Research*, *6*(3), e27.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (pp. 491–502).

Ullmann, S. (1964). *Language and style: collected papers* (Vol. 1). B. Blackwell.

Vechtomova, O., & Robertson, S. (2014). Integration of Collocation Statistics into the Probabilistic Retrieval Model. In *22nd Annual Colloquium on Information Retrieval Research* (pp. 165–177).

Xu, J., & Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems (TOIS)*, *18*(1), 79–112.

Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2* (pp. 454–460).