Paraphrase Detection based on Vector Space Model: A Study of Utilization of Semantic Network for Improving Information

Nurwati,Yudi Santoso,Krisna Adiyarta
Universitas Budiluhur

Keywords: Paraphrasing, Vector Space Model, Precision and Recall

Abstract: Paraphrasing if seen in plain view, does not look like it, so we need a technique or model that can measure the level of similarity between documents that will compare these documents. Vector Space Model is a standard approach model used to find similarities between documents. This study aims to find a system model that can be applied to paraphrase detection applications by utilizing semantic information as a tool that is integrated with the Vector Space Model. This study will use prototyping research strategies. The approach taken in conducting this investigation is to compare the performance of the system prototype developed according to the research hypothesis with a standard prototype that is built according to a standard framework. In its investigation, this research will use Confusion matrix as the most popular tool in evaluating system performance using accuracy performance criteria, namely Precision and Recall. In this way, it is expected that the semantic network model, data structure model, and algorithm that can be integrated with the vector space model to produce a paraphrase detection system that has a perfect performance is expected.

1 INTRODUCTION

Paraphrasing is a linguistic term which means re-expressing a concept in another way in the same language, but without changing its meaning. Paraphrasing gives the author the possibility to emphasize somewhat differently from the original author.

The rapid development of technology makes it easy for information users to find and find the information needed. The internet is one of the most widely used technology products for information seekers by providing sources of information from the authors themselves as well as duplicating with or without including the original authors. The ease of getting information and documents through internet media creates new problems because it turns out that documents are still found without mentioning the source of the document's author. It was not known intentionally or accidentally. Posts or pieces of writing taken from other people's writings, intentionally or unintentionally, if not correctly and adequately referenced, can be categorized as plagiarism, according to (Isa et al. 2014).

The task of identifying paraphrases has become mainstream in the research area for natural language processing. The success of application development that utilizes semantic similarities is very dependent on the ability of the system (algorithm) to determine whether or not there is a semantic relationship between two words or terms. In this study we see the problem of paraphrasing in two forms, say A and B, it is viewed as semantic quantification, the relationship between two texts, for example, to what extent text A has the same meaning as text B (paraphrase relationship) or the extent of text A part of semantic text B (entailment relationship). Considering this fact, the formulation of the problem from this study is It is difficult to build a perfect paraphrase detection system that can detect text by considering the physics of the two texts.

This study focuses on the use of semantic network, which is used as a tool to represent knowledge. This study focuses on the use of semantic networks which are used as a tool to represent paraphrase knowledge as a tool to improve the performance of paraphrase detection systems. Then the results of the vector space model with the semantic network are expected to produce a paraphrase detection system that is better than the existing one.

Nurwati, ., Santoso, Y. and Adiyarta, K.
Paraphrase Detection based on Vector Space Model: A Study of Utilization of Semantic Network for Improving Information.
DOI: 10.5220/0008931301720176
In *Proceedings of the 1st International Conference on IT, Communication and Technology for Better Life (ICT4BL 2019)*, pages 172-176
ISBN: 978-989-758-429-9

2. METHODOLOGY

In this detection, paper paraphrase uses a vector space model-based method. The initial stage describes the initial work in terms of conceptual aspects. This includes conducting a literature review related to research fields such as the definition of a text similarity detection system and Semantic Network. The initial stage describes the initial work in terms of conceptual aspects. This includes conducting a literature review related to research fields such as the definition of a text similarity detection system and Semantic Network. Some of the processes that go through are the tokenization process (this process has gone through a case-folding process), then the filtering process is continued. The process is continued with the stemming process to get the basic syllables using the Porter algorithm. The next step is the VSM (Vector Space Model) Construction process then compares the similarity of meaning (semantic measurement). The general architecture of the proposed method is shown in Figure 1.

The data tested is the Microsoft Research (MSR) paraphrase data set taken in Dolan et al. 2004 was quoted from (Daniel Ramage, Anna N. Rafferty 2009). Microsoft Paraphrase (MSR) paraphrase is a collection of 5801 pairs of sentences collected automatically from newswire over 18 months. Each pair is written by two judges with their binary / not one of the sentence sentences that are valid from the other. Annotator is asked for the value of each sentence pair quite accordingly. The Interannotator Agreement is 83%. However, 67% of the couples succeeded in paraphrasing, as well as the literature did not reflect the scarcity of paraphrasing. Data sets are present in pre-split to 4076 pair training and 1725 test pairs. Mihalcea et al. 2006 quoted from (Daniel Ramage, Anna N. Rafferty 2009) give results for several steps that underlie lexical-semantic linkages. This is further divided into action-based actions (using Latent Semantic Analysis based on Landauer et al., 1998 quoted from (Daniel Ramage, Anna N. Rafferty 2009) and exchange of knowledge and knowledge-based resources driven by WordNet. In this unsupervised experimental setting, we consider using only the threshold similarity values of our system and from the Mihalcea algorithm to determine paraphrasing or non-paraphrase considerations. For consistency with research we

use a threshold of 0.5 to measure whether the pair of sentences tested have the same meaning or not.

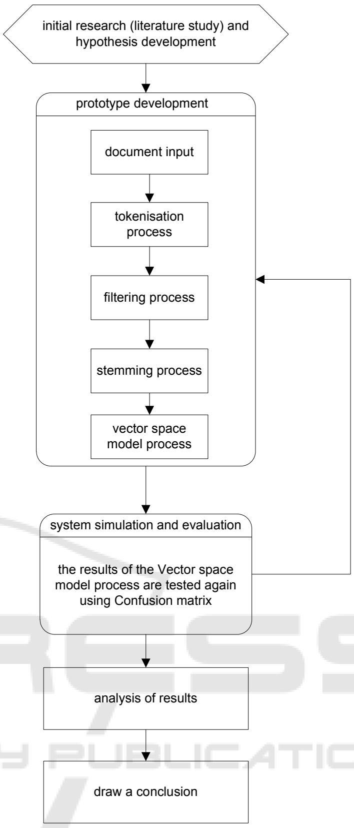
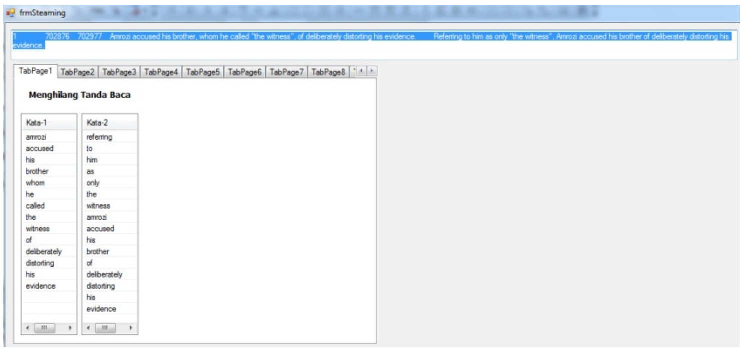

initial research (literature study) and hypothesis development
prototype development
document input
tokenisation process
filtering process
stemming process
vector space model process
system simulation and evaluation
the results of the Vector space model process are tested again using Confusion matrix
analysis of results
draw a conclusion

Figure 1 The general architecture of the proposed method

3. RESULTS

a. Experiment Standard Approach
Pre-processing
Tokenizing or tokenization is cutting each word in a sentence or parsing by using a space as a delimiter that will produce a word token. In the tokenization stage, it also removes punctuation or special characters (Bania Amburika, Yulison Herry Chrisnanto 2016).
Figure 2 to display the process.



1 702876 702977 Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

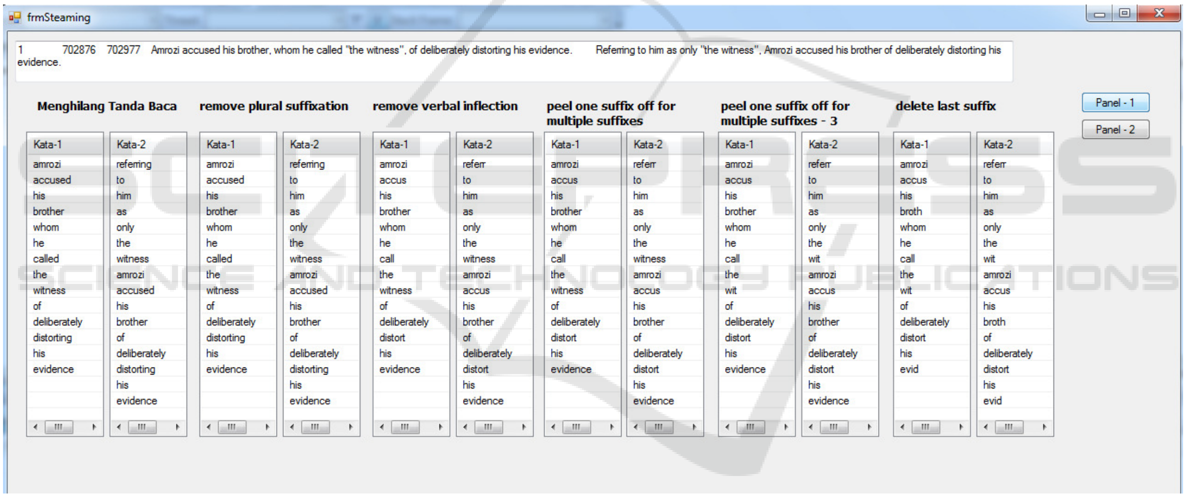
Menghilang Tanda Baca

Kata-1	Kata-2
amrozi	referring
accused	to
his	him
brother	as
whom	only
he	the
called	witness
the	amrozi
witness	accused
of	his
deliberately	brother
distorting	of
his	deliberately
evidence	distorting
	his
	evidence

Figure 2 Tokenisation process

Stemming Process

Stemming is returning words obtained from the filtering results to its basic form, eliminating the initial affixes (prefix) and final affixes (suffix) so that the necessary words (Bania Amburika, Yulison Herry Chrisnanto 2016) are obtained. For this stage of stemming, we use the stemming of the Porter algorithm. In figure 4 to show the process.



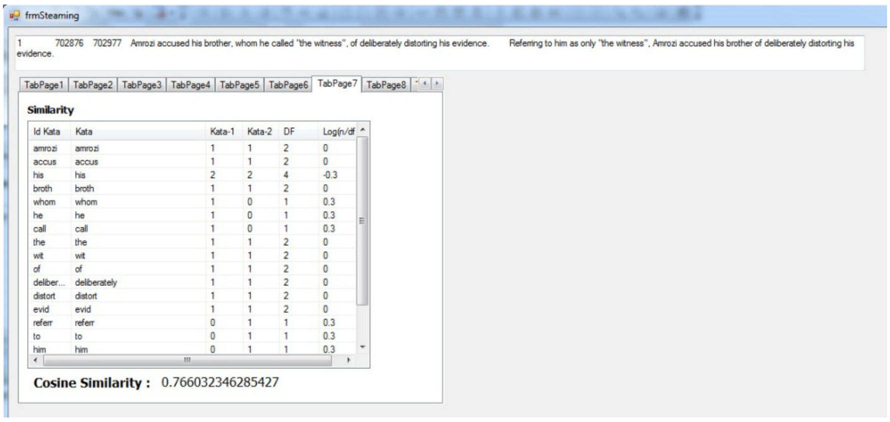
1 702876 702977 Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

Menghilang Tanda Baca		remove plural suffixation		remove verbal inflection		peel one suffix off for multiple suffixes		peel one suffix off for multiple suffixes - 3		delete last suffix	
Kata-1	Kata-2	Kata-1	Kata-2	Kata-1	Kata-2	Kata-1	Kata-2	Kata-1	Kata-2	Kata-1	Kata-2
amrozi	referring	amrozi	referring	amrozi	referr	amrozi	referr	amrozi	referr	amrozi	referr
accused	to	accused	to	accus	to	accus	to	accus	to	accus	to
his	him	his	him	his	him	his	him	his	him	his	him
brother	as	brother	as	brother	as	brother	as	brother	as	broth	as
whom	only	whom	only	whom	only	whom	only	whom	only	whom	only
he	the	he	the	he	the	he	the	he	the	he	the
called	witness	called	witness	call	witness	call	witness	call	wit	call	wit
the	amrozi	the	amrozi	the	amrozi	the	amrozi	the	amrozi	the	amrozi
witness	accused	witness	accused	witness	accus	witness	accus	wit	accus	wit	accus
of	his	of	his	of	his	of	his	of	his	of	his
deliberately	brother	deliberately	brother	deliberately	brother	deliberately	brother	deliberately	brother	deliberately	broth
distorting	of	distorting	of	distort	of	distort	of	distort	of	distort	of
his	deliberately	his	deliberately	his	deliberately	his	deliberately	his	deliberately	his	deliberately
evidence	distorting	evidence	distorting	evidence	distort	evidence	distort	evidence	distort	evid	distort
	his		his		his		his		his		his
	evidence		evidence		evidence		evidence		evidence		evid

Figure 4 Stemming Process

b. TFIDF

calculations and similarity calculations between documents using a vector space model. TF is a simple weighting where it is crucial whether or not a word is assumed to be proportional to the number of occurrences of the word in the document, while IDF is a weighting that measures how important a word is in a document when viewed globally in all documents (Bania Amburika, Yulison Herry Chrisnanto 2016). After going through the TFIDF calculation, proceed with the vector space model process in figure 5. Similar results cosine similarity use vector space model 0.766.



1 702876 702977 Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

Similarity

Id Kata	Kata	Kata-1	Kata-2	DF	Log(n/df
amrozi	amrozi	1	1	2	0
accus	accus	1	1	2	0
his	his	2	2	4	-0.3
broth	broth	1	1	2	0
whom	whom	1	0	1	0.3
he	he	1	0	1	0.3
call	call	1	0	1	0.3
the	the	1	1	2	0
wit	wit	1	1	2	0
of	of	1	1	2	0
deliber...	deliberately	1	1	2	0
distort	distort	1	1	2	0
evid	evid	1	1	2	0
referr	referr	0	1	1	0.3
to	to	0	1	1	0.3
him	him	0	1	1	0.3

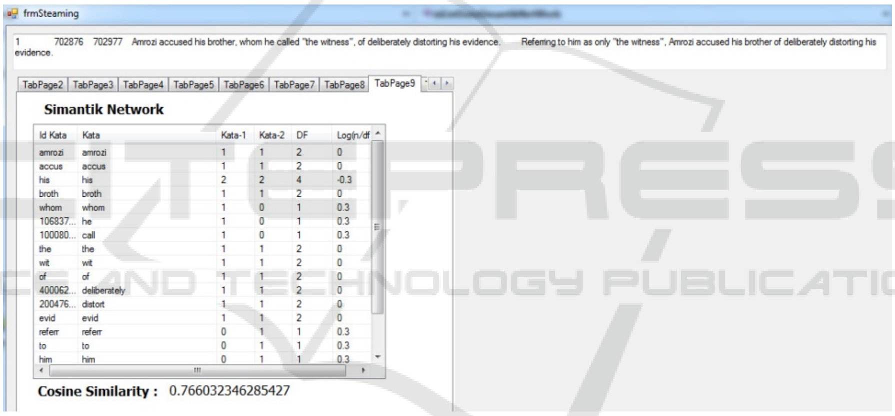
Cosine Similarity : 0.766032346285427

Figure 5 Similar results with the VSM formula

c. Similarity Results with the Semantic Network method

The results of similarities use the Vector Space Model with the Semantic Network method in .

figure 6. Similar cosine similarity uses semantic network method 0.766



1 702876 702977 Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

Simantik Network

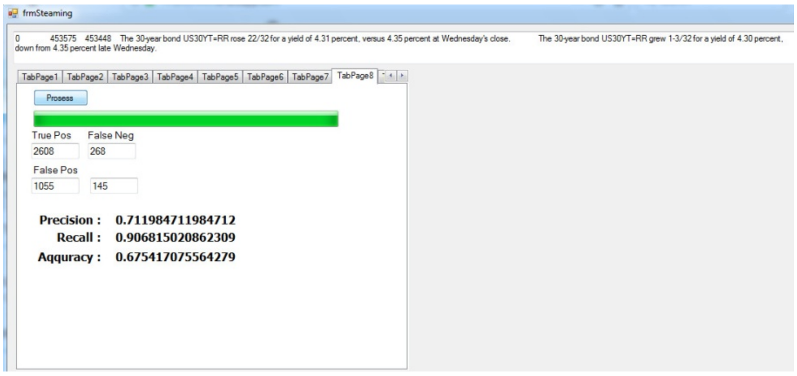
Id Kata	Kata	Kata-1	Kata-2	DF	Log(n/df
amrozi	amrozi	1	1	2	0
accus	accus	1	1	2	0
his	his	2	2	4	-0.3
broth	broth	1	1	2	0
whom	whom	1	0	1	0.3
106837...	he	1	0	1	0.3
100080...	call	1	0	1	0.3
the	the	1	1	2	0
wit	wit	1	1	2	0
of	of	1	1	2	0
400062...	deliberately	1	1	2	0
200476...	distort	1	1	2	0
evid	evid	1	1	2	0
referr	referr	0	1	1	0.3
to	to	0	1	1	0.3
him	him	0	1	1	0.3

Cosine Similarity : 0.766032346285427

Figure 6 Similar results with the Semantic Network method

d. Confusion Matrix table

Next is the confusion matrix table for the similarity of sentences using Semantic Network figure 7.



0 453575 453448 The 30-year bond US30YT=RR rose 22/32 for a yield of 4.31 percent, versus 4.35 percent at Wednesday's close. The 30-year bond US30YT=RR grew 1-3/32 for a yield of 4.30 percent, down from 4.35 percent late Wednesday.

True Pos	False Neg
2608	268
False Pos	
1055	145

Precision : **0.711984711984712**
Recall : **0.906815020862309**
Aqqur,acy : **0.675417075564279**

Figure 7 confusion matrix table for the similarity of sentences using a semantic network

4. CONCLUSIONS AND DISCUSSION

The results and analysis of word similarity by comparing the Vector Space Model method with Semantic Network in table 1.

Table 1 comparing the Vector Space Model method with Semantic Network

Method	Cosine Similarity
Vector Space Model	0,766
Semantic Network	0,766

From the table above, the Vector Space Model and Semantic Network methods have the same value. This is because the words in the Microsoft Research Paraphrase Corpus (MSRP) data as standard datasets are not all available in the Prolog dataset (datasets containing words that have similar meanings). In the prologue data set, if there is a word id, there is a meaningful code in the MSRP dataset, but it turns out that many do not have the code. If there is no meaning code in the prologue dataset (the similarity of the meaning of the word), there is no word found. Results and analysis of accurate measurement of precision and recall Vector Space Models with Semantic Network at table 2.

Table 2. Results and analysis of accurate measurement of precision and recall Vector Space Models with Semantic Network

Method	Precision	Recall	Accuracy
Vector Space Model	0,71	0,90	0,67
Semantic Network	0,71	0,91	0,67

The value of precision or measurement value of the quality of how useful the search system is in both the Vector Space Model and Semantic Network measurements is equally worth 0.71. This means that both in the Vector Space Model and Semantic Network method the ability to recall information is useful because it is 71%.

The recall value of the quality of how complete the search system displays the relevant results in either the Vector Space Model is 0.90, and the Semantic Network is 0.91. This means that both in the Vector Space Model and Semantic Network method the ability to recalculate the acquisition value is challenging to measure because the number of all relevant documents in the database is huge, which is around 90%.

Accuracy values are defined as the level of closeness between predictive values , and the actual value of Space Space Model and Semantic Network are both worth 0.67 or 67%.

REFERENCES

Bania Amburika, Yulison Herry Chrisnanto, W.U., 2016. Teknik Vector Space Model (VSM) Dalam Penentuan Penanganan Dampak Game Online Pada Anak. publikasiilmiah.unwahas.ac.id. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjuyNeFyf7gAhXu4nMBHbk0B-wQFjAAegQIBBAC&url=https://publikasiilmiah.unwahas.ac.id/index.php/PROSIDING_SNST_FT/article/viewFile/1512/1595&usg=AOvVaw0.

Daniel Ramage, Anna N. Rafferty, and C.D.M., 2009. Random Walks for Text Semantic Similarity. Proceeding TextGraphs-4 Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, pp.23–31. Available at: https://dl.acm.org/citation.cfm?id=1708131.

Isa, T.M. et al., 2014. Mengukur Tingkat Kesamaan Paragraf Menggunakan Vector Space Model untuk Mendeteksi Plagiarisme. Available at: https://rp2u.unsyiah.ac.idindex.php/welcome/prosesDownload/588/5.