



**CALCULATION OF TEXTUAL SIMILARITY USING SEMANTIC
RELATEDNESS FUNCTION**

AMMAR RIADH KAIRALDEEN

DECEMBER / 2014

**CALCULATION OF TEXTUAL SIMILARITY USING SEMANTIC
RELATEDNESS FUNCTION**

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY**

**BY
AMMAR RIADH KAIRALDEEN**

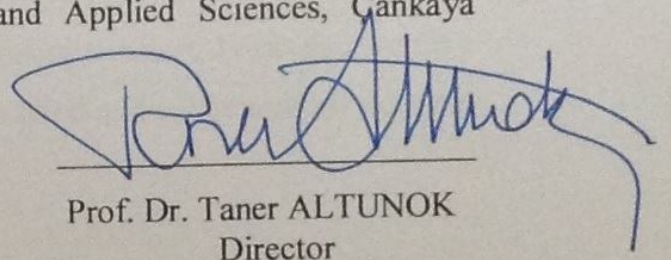
**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF
COMPUTER ENGINEERING**

DECEMBER / 2014

Title of the Thesis: **Calculation of Textual Similarity Using Semantic Relatedness Function**

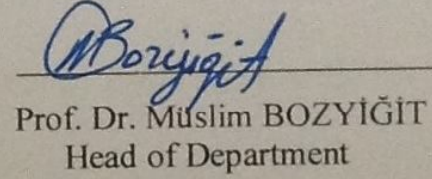
Submitted by **Ammar Riadh KAIRALDEEN**

Approval of the Graduate School of Natural and Applied Sciences, Cankaya University.



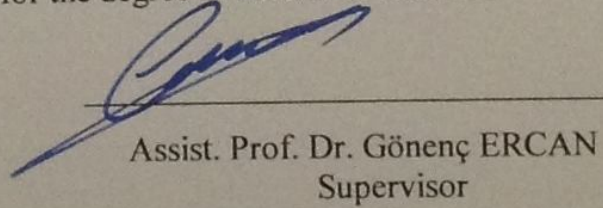
Prof. Dr. Taner ALTUNOK
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.



Prof. Dr. Müslim BOZYIĞIT
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.



Assist. Prof. Dr. Gönenç ERCAN
Supervisor

Examination Date: 29.12.2014

Examining Committee Members

Assist. Prof. Dr. Gönenç ERCAN

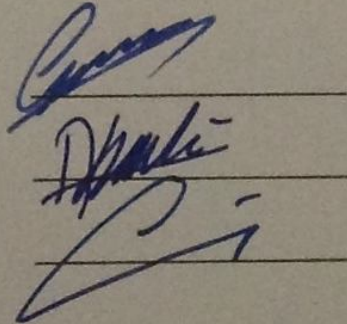
(Çankaya Univ.)

Assist. Prof. Dr. Abdulkadir GÖRÜR

(Çankaya Univ.)


Assoc. Prof. Dr. Ersin Elbaşı

(İpek Univ.)



STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Ammar, Riadh KAIRALDEEN
Signature : 
Date : 29 . 12 . 2014

ABSTRACT

CALCULATION OF TEXTUAL SIMILARITY USING SEMANTIC RELATEDNESS FUNCTION

KAIRALDEEN, Ammar Riadh

M.Sc., Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Gönenç ERCAN

December 2014, 40 Pages

Finding the similarity between two sentences is an essential task in different fields such as natural language processing (NLP) and information retrieval (IR). Semantic relatedness similarity between two sentences is concerned with measuring how two sentences share the same meaning. Over the last decade, different methods for measuring sentence similarity have been proposed in the literature. Some methods use word semantic relatedness function in sentence similarity calculations. This thesis aims to compare these methods using four data sets selected from different fields, providing a testable of a various range of writing expressions to challenge the selected methods. Results show that the use of corpus-based word semantic similarity function has significantly outperformed that of WordNet-based word semantic similarity function in sentence similarity methods. Moreover, we propose a new sentence similarity measure method by extending an existing method in the literature called Overall similarity. Furthermore, the results show that the proposed method has significantly improved the performance of the Overall method. All the selected methods are tested and compared with other state-of-the-art methods.

Keywords: Information Retrieval, Semantic Similarity, Natural Language Semantics.

ÖZ

SEMANTİK İLİŞKİ FONKSİYONUNU KULLANARAK METİN BENZERLİKLERİNİN HESAPLANMASI

KAIRALDEEN, Ammar Riadh

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi: Yrd. Doç. Dr. Gönenç ERCAN

Aralık 2014, 40 Sayfa

İki cümle arasındaki benzerliklerin bulunması,(NLP) Doğal Dil İşleme ve (IR) Bilgi Alma gibi değişik alanlarda önemli bir görevdir. Semantik (Anlamsal) Benzerlik iki cümlenin nasıl aynı anlamları paylaştığının ölçülmesiyle ilgilidir. Son 10 yıl içerisinde, değişik cümle benzerlik ölçüm yöntemleri literatürde önerilmiştir. Bazı yöntemler cümle benzerlik hesaplamalarında Kelimenin Semantik Benzerliği işlevini kullanmaktadır. Bu tez, farklı alanlardan seçilen dört veri setini kullanarak seçilen yöntemlerle karşılaştırılabilecek test edilebilir ve çeşitli aralıklardaki yazım ifadelerini sağlamayı ve bu yöntemleri karşılaştırmayı amaçlar. Sonuçlar kelime benzerlik yöntemleri içerisinde Corpus-tabanlı Kelime Benzerlik işlevinin WordNet-tabanlı Kelime Semantik Benzerlik işlemine göre daha iyi bir performans çıkardığını gösterir. Buna ek olarak, literatürde mevcut olan Overall Similarity yöntemi genişletilerek yeni kelime benzerlik ölçüm yöntemi önerilmiştir. Ayrıca, sonuçları önerilen bu yeni yöntem, mevcut olan Overall Similarity yönteminin performansını arttırmıştır.Böylece seçilmiş tüm yöntemler test edilmiş ve diğer en son teknolojiler ile karşılaştırılmıştır..

Anahtar Kelimeler: Bilgi Alma, Semantik Benzerlik, Doğal Dil Semantikleri.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and deep appreciation to Dr. Gönenç for his unlimited guidance, assistance, encouragement and tremendous patience throughout this research. Besides my advisor, I would like to thank the rest of my thesis committee for their constructive criticism and insight.

I am deeply and forever indebted to my parents for their love, support and encouragement throughout my entire life, I would like to thank my wife Zahraa and my sweet daughters Safa and Sama for their unconditional support, friendship and love through good times and bad. Finally, I thank my siblings who encouraged me with their best wishes and moral support.

TABLE OF CONTENTS

STATEMENT OF NON PLAGIARISM.....	iii
ABSTRACT.....	iv
ÖZ.....	v
ACKNOWLEDGEMENTS.....	vi
TABLE OF CONTENTS.....	vii
LIST OF FIGURES.....	viii
LIST OF TABLES.....	x
LIST OF ABBREVIATIONS.....	xi

CHAPTERS:

1. INTRODUCTION	
1.1. Semantic Textual Similarity (STS).....	1
1.2. Objectives.....	1
1.3. Outlines	2
2. THEORETICAL FOUNDATION	
2.1. Definitions	3
2.1.1. Lexical relations.....	3
2.1.1.1. Synonymy and hyponymy	4
2.1.1.2. Polysemy	4
2.1.2. Lexical resources	3
2.1.2.1. Roget's thesaurus	5
2.1.2.2. Controlled vocabularies	5
2.1.2.3. WordNet	5
2.2. Applications	6
2.2.1. Information retrieval	6
2.2.2. Natural language processing	7
2.3. Computing Lexical Similarity	7

2.3.1.	Computing word to word lexical similarity.....	7
2.3.1.1.	WordNet based semantic relatedness similarity functions	8
2.3.1.1.1.	The Leacock & Chodorow (LCH) function	9
2.3.1.1.2.	The Wu & Palmer (WUP) function	10
2.3.1.1.3.	The Resnik (RES) function	10
2.3.1.1.4.	The LIN function	11
2.3.1.1.5.	The Jiang & Conrath (JCN) function	12
2.3.1.1.6.	The Lesk function	12
2.3.1.1.7.	Hirst and St-Onge (HSO) function	13
2.3.1.2.	Corpus-based similarity method	13
2.3.1.2.1.	Latent Semantic Analysis (LSA) function	14
2.3.1.2.2.	Pointwise Mutual Information (PMI-IR) function..	14
2.3.2.	Computing sentence to sentence similarity	15
2.3.2.1.	Cosine similarity method	15
2.3.2.2.	Dissimilarity significance methods	15
2.3.2.3.	Text canonicalization methods	16
2.3.2.4.	Semantic similarity methods	16
2.3.2.5.	A semantic similarity approach to paraphrase	17
2.3.2.6.	Word order similarity	17
2.3.2.7.	Overall sentence similarity	17
3.	EVALUATION METHODOLOGY	
3.1	Measures	19
3.1.1.	Pearson correlation	19
3.1.2.	Test of difference between two correlations	20
3.2.	Data Sets	21
3.2.1.	Corpus	21
3.2.2.	Document preprocessing	23
3.2.2.1.	Tokenization	23
3.2.2.2.	Normalization	23

3.2.2.3.	Stop words	23
3.2.3.	Vector space model	23
3.3.	Computing Similarity	23
3.3.1.	Cosine similarity method	24
3.3.2.	Semantic matrix method	24
3.3.3.	Word order method	25
3.3.4.	Overall sentence similarity method	26
3.3.5.	Enhanced overall sentence similarity method	26
4.	RESULTS AND DISCUSSION	
4.1.	Experiments	29
4.2.	Compare Results	29
4.2.1.	Comparison of lexical semantic relatedness with cosine similarity	29
4.2.2.	Overall similarity and enhanced overall similarity	31
4.2.3.	All corpuses comparisons and rank the results	32
4.2.4.	Compare with other methods uses the same data sets	36
5.	CONCLUSIONS AND FUTURE WORK	39
5.1.	Conclusions	39
5.2.	Future Work	40
	REFERENCES	R1
	APPENDICES	A1
A.	CURRICULUM VITAE	A1

LIST OF FIGURES

FIGURES

Figure 1	Hypernyms as a type of lexical relation	3
Figure 2	Semantic relations in WordNet	6
Figure 3	Corpus-based and WordNet-based used functions	8
Figure 4	WUP metric paths and distance	10
Figure 5	WordNet taxonomy represent IS-A links adapted for Resnik	11
Figure 6	Paraphrasing & Non-paraphrasing example using Qiu et al.	16

LIST OF TABLES

TABLES

Table 1	Pearson Correlation Values Range	20
Table 2	Number of Sentences, Words and Unique Words in the Corpora.	22
Table 3	Cosine Similarity Calculation Example	24
Table 4	Word Matrix Score for the Two Sentences, Using LSA Metric.....	25
Table 5	Sentence 1 and 2 Vectors	25
Table 6	Joint Word Set.....	26
Table 7	Word Order Example.	26
Table 8	$S_{overall}$ Between two Sentences	26
Table 9	$SE_{overall}$ Between two Sentences	27
Table 10	$S_{overall}$ and $SE_{overall}$ Between two Sentences.....	27
Table 11	Overall Similarity and Enhanced Overall Similarity Examples.	28
Table 12	Pearson’s Correlations for Cosine Similarity and Semantic Matric.	29
Table 13	Pearson’s Correlations for all Corpuses for Cosine Similarity and Semantic Metrics	30
Table 14	Pearson’s Correlations to Overall Similarity and Enhanced Overall Similarity.	31
Table 15	Pearson’s Ccorrelations for all Corpuses Using Overall Similarity and Enhanced Overall Semantic Similarity.	32
Table 16	Headlines Pearson Correlation Rank for all Methods.....	33
Table 17	Images Pearson Correlation Rank for all Methods.	34
Table 18	OnWN Pearson Correlation Rank for all Methods.....	34
Table 19	MSRvid Pearson Correlation Rank for all Methods.	35
Table 20	Pearson Correlation Rank for all Corpuses.	36
Table 21	Compare with other Results that use the Same Data Sets.....	37
Table 22	Results Rank with Other Researchers	37

LIST OF ABBREVIATIONS

SR	Semantic Relatedness
STS	Semantic Textual Similarity
NLP	Natural Language Processing
IR	Information Retrieval
MeSH	Medical Subject Headings
ERIC	Educational Resources Information Centre
LCSH	Library of Congress Subject Headings
VSM	Vector Space Model
AI	Artificial Intelligence
TOEFL	Test of English as a Foreign Language
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
PMI-IR	Pointwise Mutual Information - Information Retrieval
PPMC	Pearson Product Moment Correlation
LCS	Longest Common Subsequence
LCH	Leacock & Chodorow
WUP	Wu & Palmer
RES	Resnik
LIN	Lin, Dekang
JCN	Jay & Conrath
HSO	Hirst and St-Onge
SemEval	Semantic Evaluation

CHAPTER 1

INTRODUCTION

1.1. Semantic Textual Similarity (STS)

In natural language processing (NLP), determining the similarity between two sentences is a crucial task due to unexpected expressions in natural language (i.e. various range of writing expressions), and has impact in research fields that seeks to bridge the gap between NLP and different applications.

The techniques used in detecting the similarity between two long texts are different than those used for short texts. Long text techniques rely on analyzing the shared words between two texts, which cannot be used in the short text techniques where shared words can be rare or even an empty set. Thus, semantic information should be taken into account by extracting the syntactic and semantic similarity for the similarity detection techniques.

The techniques to measure the semantic similarity have been applied and developed in different fields [1, 2]. For instance, in information retrieval (IR) to solve the problem of measuring the similarity to assign a ranking score between a query and texts in a corpus [3]. In text summarization, sentence semantic similarity is used to cluster similar sentences [4]. In web page retrieval, sentence similarity can be effectively enhanced by calculating the page title similarity [5]. These are only a few examples of the applications of sentence semantic similarity. Therefore, it is important to ongoing research and development to improve this success over a wide range of applications by using similarity measures and lexical semantic resources.

1.2. Objectives

The main objective of this research is the evaluation between the methods that are based on word semantic relatedness functions and the methods that are not.

Moreover, these methods are compared using multiple data sets from different domains. Specifically, this research covers the following points:

- Comparing different sentence similarity methods that are used for paraphrase detection problem.
- Comparison between corpus-based word semantic relatedness with WordNet-based.
- Enhancing the Overall similarity method [6] that considers the word order similarity, by adding the word semantic similarity function.

1.3. Outlines

The remainder of this document is structured as follows:

Chapter Two, illustrates the theoretical foundation, this chapter is divided into three parts: first part defines terms related with STS approach and some of the lexical resources. Second part addresses some applications that utilize STS, and the third part elaborates on the theoretical foundation used in calculating the STS.

Chapter Three, describes the experiments. This chapter is divided into three parts: first part covers the measurements foundation to evaluate the results, second part describes the data sets used, and preprocessing steps. The third part illustrates how similarity is computed in our work in detail with examples.

Chapter Four, discusses the obtained results and compares the selected methods. Comparisons are based on Pearson correlation between the used methods and the human judge scores. Also comparison with other state-of-the-art research is included in this chapter.

Chapter Five, conclusions and future works, reaffirms the thesis objectives and reaches the final judgment that is obtained from the analyzed results gained from the implementation and the statistical analyzing, as well as forecasting future trends, and the need for further research to extending the scope of this thesis, some directions are presented.

CHAPTER 2

THEORETICAL FOUNDATION

2.1. Definitions

This section explains the lexical relation types between the words to illustrate the importance of measuring the similarity between the words. Also some lexical resources repositories are viewed.

2.1.1. Lexical relations

In the language, the relations between the words contribute to the reader's understanding. The author clearly knows the semantic relations among the intended sense. Related words may join together to form larger groups of related words that can extend freely over sentence boundaries.

The most common word relation types are synonyms and antonyms. There is synonym between two words if they share a meaning, antonym means the two words have opposite meaning. Hypernymy is defined as a "type-of" relation with its hypernym.

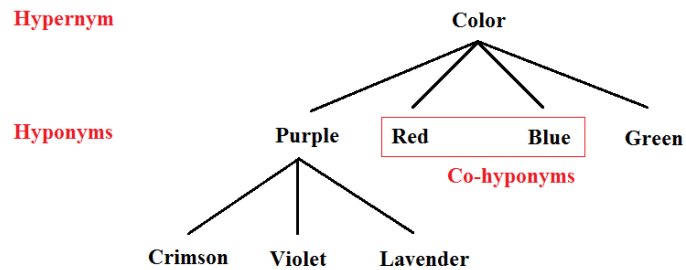


Figure 1 Hypernyms as a type of lexical relation

2.1.1.1. Synonymy and hyponymy

When words share the similar/same meaning then they are called synonyms. Humans can recognize synonymous words easily. Semantic relatedness is used as an important criteria for detecting synonymy [7].

There are three types of synonyms. The first one when the words are identical in meaning and when these words can be interchanged in the syntactic contexts, this condition is called absolute synonymy [8] or full [7]. (e.g. internet and cyberspace). The second type of synonyms is called cognitive synonymy [8] or sense synonymy [7] that is when we have two words with one sense where they are the same, but these words are different in the other sense of expression (e.g. Father and Daddy). The third type of synonyms is called plesionymy or near-synonymy [9]. It is exactly the same, the senses are very similar [7]. They are distinguished from cognitive synonyms by the fact, (e.g. foggy and misty, or freedom fighter and terrorist).

Words can be classified under these typologies of synonyms. In the end, potentially, we have to accept the fact that some words could be more suitable synonyms than others [7]. Furthermore, when a specific meaning of a word is contained in another more general word, like a semantic relationship inclusion, then it is called as hyponymy. The study of hyponymous theory is developed early in the branch of linguistics [10].

2.1.1.2. Polysemy

When words or phrases have multiple distinct senses then they are called *polysemy*. (e.g. understand and I got it) Placing words into different places in the hierarchy in the WordNet and Roget made a distinct meaning that depends on two terms similarity.

There have been several computational attempts to reduce the number of sense distinctions and increase the size of each synset in WordNet [11]. Also, another major problem is related to domain dependency of synonymy, since some words have a high similarity in a specific domain but not in others.

The semantic relations in textual similarity are acquired by extracting the knowledge between two sentences. For instance, describing a drink (e.g. “milk”) and the

quantity of this drink (e.g. “a bottle of milk”). Acquiring this knowledge is by reducing the redundancy in words and keeping the meaning intact. Finally, many applications would benefit from topical semantic similarity.

2.1.2. Lexical resources

Lexical resources are used in different applications like machine translation, information retrieval, information extraction and text summarization. Lexical knowledge obtained from different resources is used in these different applications. Below, some of these resources are listed.

2.1.2.1. Roget’s thesaurus

Roget’s Thesaurus is a system of classification. The name (Thesaurus) comes from Greek etymology (*storehouse* or *treasure*) and represents a vast treasure house for the English phrases and words.

In Roget, English words collection has a combination of idioms, arranged according to the expressed ideas. For this reason, it can be described as a reverse dictionary because it does not follow the alphabetical order of the dictionary [12]. It also includes semantically similar phrases and words, divided into verbs, adverbs, nouns, adjectives and interjections.

2.1.2.2. Controlled vocabularies

Controlled Vocabularies is thesauri giving description for every concept in a specific domain. It uses subject search as organized system of knowledge that includes terms selection that correspond to the interest topic and it retrieves all indexed documents by those terms. (e.g. *Medical Subject Headings* thesaurus MeSH [13], *Educational Resources Information Centre* thesaurus ERIC [14], and *Library of Congress Subject Headings* LCSH) [15].

2.1.2.3. WordNet

WordNet is a combination of both dictionaries and thesauri. It is inspired by the current psycholinguistic theories of the human lexical memory. WordNet consists of English nouns, verbs, adverbs and adjectives organized into synsets in which various lexical-semantic relations connect synsets together. The noun and verb synsets are

organized into hierarchies based on the hypernymy relation [16]. WordNet focuses on meanings of words instead of the word forms. The semantic relation between the synsets depends on the grammatical category, as can be seen in Figure 2 [17].

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	Noun Verb Adj Adv	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Adj Adv Noun Verb	wet, dry rapidly, slowly top, bottom rise, fall
Hyponymy (subordinate)	Noun	sugar maple, maple maple, tree tree, plant
Meronymy (part)	Noun	brim, hat gin, martini ship, fleet
Troponymy (manner)	Verb	march, walk whisper, speak
Entailment	Verb	drive, ride divorce, marry
Derivation	Adj Adv	magnetic, magnetism simply, simple

Figure 2 Semantic relations in WordNet.

2.2. Applications

This section describes some NLP, and IR applications, which benefit from semantic textual similarity.

2.2.1. Information retrieval

Lexical semantic resources are used in IR to bridge the gap between the user's information need and the relevant information resources. The problem of retrieving information overload or non-relevant information of the user need is caused by using the wrong search method. Some IR models are described below.

The first model is Boolean model [18] that is based on Boolean algebra and set theory. Here, queries are specified as conjunctions "AND", disjunctions "OR", or

negations “NOT” of index terms. The second model is the vector space model (VSM) [19]. Unlike the Boolean model, it incorporates weights for each term and supports partial matching by inspecting the relevancy between the document and the query. The Statistic Language Model [20] that is based on statistics and probability, in the beginning creates the language model for each document then evaluates the documents based on the probability of generating the query. The fuzzy set model [20] defines each query term as a fuzzy set and each document in this set has a degree of membership. However, all models incorporate ranking based on the similarity between the query and the document.

With the Internet growth in the last decades, the IR challenge enlarged because of the information location in the dynamic and incremented volume of web pages; different tools are developed and have matured to cover this challenge.

2.2.2. Natural language processing

In NLP research thesauri, WordNet and other lexical resources are used for various applications [18, 21]. Similarity measures extracted from raw text, or calculated over lexical-semantic resources, are widely used. Some paradigms that aim to enable computer detection of the linguistic meaning are described below.

The first approach is lexical semantics and ontologies; it is the study of how the knowledge is represented within a language domain. Grammaticalization semantics approach is the process by which words are transformed to be grammatical markers to detect objects and actions (i.e. nouns and verbs). Relational semantics attempts to use the links between the concepts in the sentence. Relational semantic parsing [22, 23], and coreference resolution [24]. Logical semantics aims to produce the logical form of the sentences that is much like syntactic parsing. Deep meaning and reasoning systems [25] are associated with AI and applications like language robotics, question answering and computer dialogues [26].

2.3. Computing Lexical Similarity

This section is divided into two parts. The first part covers the methods used to find the similarity between the words used as word semantic similarity function. The second part covers the methods used to find sentence similarity.

2.3.1. Computing word to word lexical similarity

Different methods attempt to calculate word to word relatedness similarity, some methods are based on information derived from a large corpus, others are based on words relations in the WordNet. In this thesis, four different methods are selected, one is corpus-based and three others are WordNet-based, as shown in Figure 3.

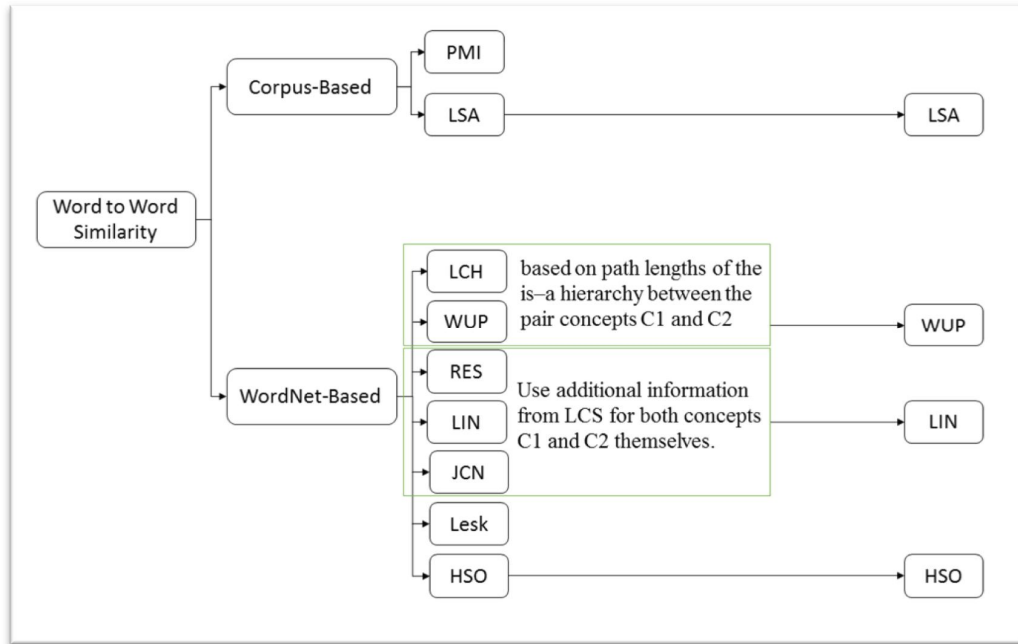


Figure 3 Corpus-based and WordNet-based used functions

2.3.1.1. WordNet based semantic relatedness similarity functions

WordNet Similarity functions give a quantitative similarity measure for two words. This is useful for the paraphrase detection task, since, if a pair of sentences shares many similar words, we can guess that this would be a good indicator they have a similar meaning as a whole.

First, clarifying some definitions is important to considering similarities between concepts or word senses, instead of the words, since words may have more than one sense. To measure similarity, we use the information in the is-a hierarchy. We consider 'car' and 'boat' to be more similar to each other than 'boat' and 'tree' since 'car' and 'boat' have a more specific common ancestor, the 'vehicle' concept.

WordNet contains separate is-a hierarchies for nouns and verbs. Similarities can only be found when both words are in one of these categories. For example, the nouns ‘dog’ and ‘cat’, and the verbs ‘run’ and ‘walk.’ Since WordNet adjectives and adverbs are not organized into is-a hierarchies, similarity measures cannot be applied. However, the concept in many ways can be related and part of each other similarity. These include part-of relationships (‘tree’ and ‘garden’), as well as opposites (‘dark’ and ‘light’) and so on.

It is possible to make additional use of relatedness measures, non-hierarchical information in WordNet, and including the sets of synonyms. Also it is possible to apply onto different concept pairs after including words with different parts of the speech. For instance “weapon” and “murder”.

The Least Common Superconcept (LCS) for the two concepts C1 and C2 is the most specific concept that is an ancestor of both C1 and C2 (e.g LCS between cat and dog is pet). In the following, word to word semantic similarity functions are used to calculate sentence similarity. Both LCH and WUP [27, 28] are based on path lengths of the is-a hierarchy between the pair concepts C1 and C2.

RES [29] use additional information from LCS by including summation of the information content for both concepts C1 and C2 themselves. The LIN [30] scales the information content of the LCS by the summation, while JCN [31] takes the difference of the calculation and the information content from LCS.

Lesk [32, 33] incorporates information from WordNet glosses, and HSO [34] classifies relations in WordNet based on the direction, and classifies relations in WordNet as having direction.

2.3.1.1.1. The Leacock & Chodorow (LCH) function

The LCH metric [27] determines the two nodes similarity by finding the path length between the nodes in the is-a hierarchy. The similarity is computed as:

$$Sim_{lch} = - \log \frac{N_p}{2D} \quad (1)$$

Where N_p is the distance between the nodes and D is the maximum depth in the is-a taxonomy.

2.3.1.1.2. The Wu & Palmer (WUP) function

The WUP metric [28] computes the nodes similarity as a function of the path length from the LCS of the nodes. The similarity between nodes C1 and C2 is:

$$sim_{wup}(C1, C2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (2)$$

Where N_1 is the number of nodes on the path from the LCS to C1, N_2 is the number of nodes on the path from the LCS to C2, and N_3 is the number of nodes on the path from the root node to the LCS. Figure 4 shows these paths and distances.

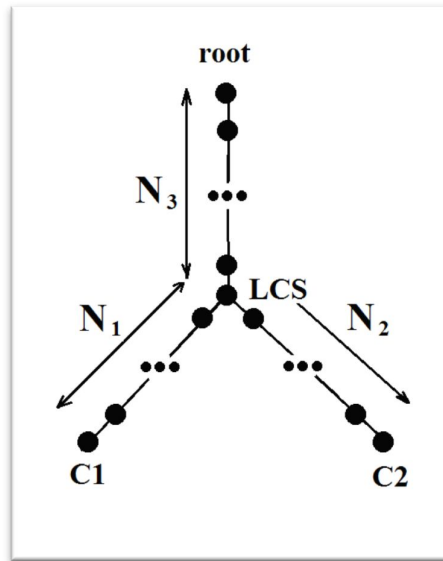


Figure 4 WUP metric paths and distance

2.3.1.1.3. The Resnik (RES) function

The RES metric [29] uses LCS information content of the two concepts. Specifically, degree of similarity between two concepts is measured by the amount of shared

information between these two concepts. This shared amount of information is indicated by the significant contents of the information of their LCS.

Formally:

$$sim_{res} = IC(LCS) \tag{3}$$

Where

$$IC(c) = -\log P(c) \tag{4}$$

And $P(c)$ is the probability of find a concept c in a large corpus.

Figure 4 below shows an example of finding the probability between two words nickel and coin, so whenever encounter the word nickel, the word coin is encountered (i.e. $p(\text{nickel}) \leq p(\text{coin})$). Particularly, if the used taxonomy has a unique top node, its probability equal to 1, so if the most-specific subsumer of a pair of concepts is the top node, their similarity equal to $-\log(1) = 0$. In the experiment to calculate P , Resnik use one-million-word Brown Corpus of American English [30].

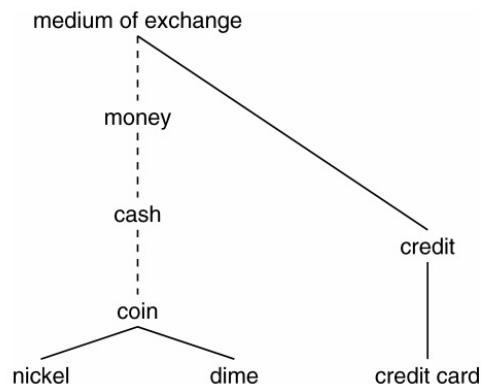


Figure 5 WordNet taxonomy represent IS-A links Adapted for Resnik

2.3.1.1.4. The LIN function

The LIN metric [31] builds on the Resnik measure by normalizing the information content of the two nodes themselves.

$$sim_{lin} = \frac{2 * IC (LCS)}{IC (N_1) + IC (N_2)} \quad (5)$$

2.3.1.1.5. The Jiang & Conrath (JCN) function

The JCN metric [32] also uses the information content idea:

$$sim_{jcn} = \frac{1}{IC (N_1) + IC (N_2) + 2 * IC (LCS)} \quad (6)$$

2.3.1.1.6. The Lesk function

In the original Lesk metric [33] the relatedness between two words is refined as the overlap of their dictionary definitions. The extended Lesk [34] uses WordNet as the dictionary to define the word. The score given to an overlap and also takes into account all concepts which are directly related to the concept via explicit relations in WordNet (hypernyms, hyponyms etc.). Lesk algorithm uses a list of relationship types called RELPAIRS to define the relationship to account reflexive in their algorithm.

$$RELPAIRS = \{(R_1, R_2) \in RELS \mid \text{if } (R_1, R_2) \in RELPAIRS \text{ then } (R_2, R_1) \in RELPAIRS\}$$

The reflexive constraint is imposed to ensure that the relatedness function is itself reflexive so that relatedness (C1,C2) = relatedness (C2,C1)

Finally, the relatedness of two synsets C1 and C2 is given by:

$$relatedness (C1, C2) = \sum_{\forall (R_1, R_2) \in RELPAIRS} score (R_1(C1), R_2 (C2)) \quad (7)$$

For example, if the set of relations

$$RELS = \{gloss, hypo, hype\}$$

And

RELPAIRS = {(gloss, gloss),(hypo, hypo),(hype, hype),(gloss, hype),(hype, gloss)}

Then:

$$\begin{aligned} relatedness(C1,C2) = & score(gloss(C1), gloss(C2)) + score(hypo(C1) + \\ & hypo(C2)) + score(hype(C1) + hype(C2)) + \\ & score(gloss(C1)+ hype(C2)) + score(hype(C1) + \\ & gloss(C2)) \end{aligned}$$

2.3.1.1.7 Hirst and St-Onge (HSO) function

The HSO metric [35] categorizes the relations between two words to three types (extra-strong, strong, and medium-strong). Then score the similarity depending on these three categorizations.

Extra-strong relation is between the word and its literal repetition, this relation type has the highest weight. Strong relation categorized also to three types: first type happens when there is a common synset to words, second type happens when there is a horizontal link (e.g. similarity, antonymy,) between synset of each word and third type happen when there is a link between a synset of each word if one word is a compound word or a phrase that includes the other. Medium-strong relation happens when a member of a set of allowable paths connects a synset of each word. The weight of a path is given by:

$$weight = C \text{ path } - length \ k * \text{ number of changes of direction} \quad (8)$$

2.3.1.2. Corpus-based semantic relatedness similarity functions

Corpus-based measures of word semantic similarity tries to identify the degree of relatedness between words using information exclusively derived from a large corpus. These approaches to word sense identification have flexibility and generality.

2.3.1.2.1. Latent Semantic Analysis (LSA) function

Is a Corpus-based measure of semantic similarity [36]. In LSA, corpus term co-occurrences are captured by dimensionality reduction means and operated by a SVD on the matrix of term-by term T for corpus representation.

The core of LSA is SVD, in linear algebra SVD is commonly used. To calculate the correlations between rows and columns in any rectangular matrix, SVD can be used. In the implementation used for this thesis, SVD decomposes the term-by-term matrix T into three matrices $T = U \Sigma_k V^T$ where Σ_k is the diagonal $k \times k$ matrix containing the k singular values of T , $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k$, and U and V are column-orthogonal matrices. When the three matrices are multiplied together, the original term-by-term matrix is re-composed. Typically k' is chosen such that $k' \ll k$, obtaining the approximation $T \cong U \Sigma_{k'} V^T$.

The U matrix contains vectors orthogonal to each other from the original rows, and the second matrix V contains orthogonal vectors from the original column values, and finally $\Sigma_{k'}$ matrix is composed of scaling values, such that when they are all multiplied together, we obtain the initial term-term matrix.

The dimensionality reduction using SVD entails the abstraction of meaning by collapsing similar contexts while discounting those that are noisy and irrelevant, hence transforming the real world word-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of the words.

The LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. Moreover, LSA benefits from using vector space model that allow representing all vectors in a multi dimension space then compare between them.

2.3.1.2.2. Pointwise Mutual Information (PMI-IR) function

PMI use collected data from Information Retrieval (PMI-IR) as an unsupervised measure for the measuring of semantic similarity of words. [37, 38]. The equation is

based on word co-occurrence counts collected by very large corpora (e.g. the web). For each word pair W_1 and W_2 , their PMI-IR are measured as:

$$PMI - IR(w_1, w_2) = \log_2 \frac{P(w_1 \& w_2)}{P(w_1) * P(w_2)} \quad (9)$$

This Equation used to measure the semantic similarity and the degree of statistical dependence between W_1 and W_2 .

Another measure [39] for document-by-term matrices and also used Entropy based weighting function [40]. The information content of the co-occurrence probability is multiplied by the Entropy of the second term.

$$A_{ij} = \log (1 + C_{ij}) \left(- \sum_k P(w_i, w_k) \log (p(w_i, w_k)) \right) \quad (10)$$

2.3.2. Computing sentence to sentence similarity method

Several methods tried to calculate the semantic similarity in short text, we try to list most famous methods.

2.3.2.1. Cosine similarity method

Cosine similarity is a popular vector based similarity measure in both information retrieval and text mining. In this approach compared strings are transformed into vector space so that the cosine of the angle between the vectors can be used to calculate the similarity [41]. This approach is often paired with other approaches (e.g. LSA) to limit the dimensionality of the vector space.

$$Similarity = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| |\vec{B}|} \quad (11)$$

2.3.2.2. Dissimilarity significance method

This approach detects dissimilarities between sentences, unlike most systems that focus on sentence similarity. After detecting dissimilarities, the method makes its paraphrase judgment based on the significance of such dissimilarities [42].

In this method two-phase processes are used, first phase identifies the key semantic content units in each sentence. These are then paired off. Second phase identifies any extraneous information nuggets remaining, and then the significance of these is judged. Then the similarity is compared using a simple lexical matching technique. If the sentences are insignificant (i.e. it contains unpaired nuggets) then a positive meaning is given.

<p>Paraphrase (+PP)</p> <p><u>Authorities said</u> a young man injured Richard Miller. Richard Miller was <i>hurt</i> by a young man</p>
<p>Non-Paraphrase (-PP)</p> <p>The technology-laced Nasdaq Composite Index .IXIC <i>added</i> 1.92 points, or 0.12 percent, at 1,647.94. The technology-located Nasdaq Composite Index .IXIC <i>dipped</i> 0.08 of a point to 1,646.</p>

Figure 6: Paraphrasing & Non-paraphrasing example using Qiu et al.

2.3.2.3. Text canonicalization method

Creating canonicalized forms of sentences is the main object in this approach. It is implemented so that texts with similar meaning are more likely to be transformed into the same surface texts than those with different meaning. Once the text is transformed in this way, simple lexical matching techniques are used to compare the transformed text [43].

2.3.2.4. Semantic similarity method

This approach tried to find the similarity of text segments T1 and T2 by using the following scoring function [44]:

$$f(x) = \frac{1}{2} \left(\frac{\sum w \in \{T1\} (\max Sim (w, T2) * idf(w))}{\sum w \in \{T1\} idf(w)} + \frac{\sum w \in \{T2\} (\max Sim (w, T1) * idf(w))}{\sum w \in \{T2\} idf(w)} \right) \quad (12)$$

Where $maxSim(w, T)$ is the maximum similarity score found between word w and words in T according to the given word-to-word similarity measure and $idf(w)$ is the inverse document frequency [45] of the word, which indicates its specificity.

2.3.2.5. Semantic similarity matrix to paraphrase method

All the similarity scores between all word pairs in the sentences are taken into account. In this approach all each sentence are represented as a binary vector (with elements equal to 1 if a word is present and 0 otherwise), \vec{a} and \vec{b} . The formula below shows how the similarity between these sentences can be computed [46].

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} W \vec{b}^T}{|\vec{a}| |\vec{b}|} \quad (13)$$

Where the W is a semantic similarity matrix containing information about the similarity of word pairs.

2.3.2.6. Word order similarity method

The information order effect the sentence meaning if the same words used into two different sentence arrangement, therefore the word order vector are composed by the entity of the words carried by the two sentences. Then word order similarity value is calculated by normalized difference of word order using the formula below [46].

$$Sr(S1, S2) = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \quad (14)$$

Where S_r is the word order similarity between two sentences S_1, S_2 that is calculated by finding the normalized differences of two vectors r_1 and r_2 of the word order set.

2.3.2.7. Overall sentence similarity method

It is a combination of both cosine similarity and structures information between two sentences to find the preferable similarity measure in one formula as below.

$$S_{overall}(S_1, S_2) = \lambda \frac{\vec{a} \cdot \vec{b}^T}{|\vec{a}| |\vec{b}|} + (1 - \lambda) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (15)$$

$$S_{overall}(S_1, S_2) = \lambda S_s + (1 - \lambda) S_r \quad (16)$$

Where S_s and S_r are explained in sections 2.3.2.5 and 2.3.2.6. And the coefficient λ is determine the related contributions of S_s and S_r of the overall similarity calculation, wherever $\lambda \leq 1$ also the sentence structure is major in the processing text, then the value of λ should be greater than 0.5, (i.e. $\lambda \in (0.5, 1]$) [47].

CHAPTER 3

EVALUATION METHODOLOGY

3.1 Measures

In order to interpret the results as simply as possible, a statistical method is applied to analyze and present the results.

3.1.1 Pearson correlation

It is the most commonly used measure in statistic for the correlation between sets of data and how well they are related. The full name is the Pearson Product Moment Correlation or PPMC. It shows the linear relationship between two sets of data and attempts to draw a line of best fit through the data of two variables, then indicates how far away all these data points are to this line of best fit. Two letters are used to represent the Pearson correlation ρ for a population and the letter r for a sample [47].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (17)$$

Where

- n = number of pairs of scores
- $\sum xy$ = sum of the products of paired scores
- $\sum x$ = sum of x scores
- $\sum y$ = sum of y scores
- $\sum x^2$ = sum of squared x scores
- $\sum y^2$ = sum of squared y scores

The Possible Values for the Pearson Correlation results will be between -1 and 1. It will very rarely be 0, -1 or 1. Mostly it is a number between these values. The closer the value of r gets to zero, the greater the variation the data points are around the line of best fit, Table 1 explains the correlation coefficient values.

Exactly +1	Perfect positive peak for linear relationship
+ 0.70	Strong positive linear relationship
+ 0.50	Moderate positive linear relationship
+ 0.30	Weak positive linear relationship
0	No linear relationship
- 0.30	Weak negative linear relationship
- 0.50	Moderate negative linear relationship
- 0.70.	Strong negative linear relationship
Exactly -1	perfect negative peak for linear relationship

Table 1 Pearson Correlation Values Range

3.1.2. Test of difference between two correlations

After calculating the correlation between two data sets, a T test is required to determine if two correlations are statistically significant or not. In an instance from our calculated result, one method obtained 0.733 by using cosine metric to find the similarity between two sentences, the correlation calculated between the cosine metric results and the human judges result. On the other hand using semantic matrix 0.708 correlation is obtained, then to calculate t-test between two correlations below formula is applied [48].

$$t = (r_{jk} - r_{jh}) \sqrt{\frac{(1 - n)(1 + r_{kh})}{2 \left(\frac{n-1}{n-3}\right) * |R| + r^2(1 - r_{kh})^3}} \quad (18)$$

Where

$$|R| = 1 - r_{jk}^2 - r_{jh}^2 - r_{kh}^2 + 2 r_{jk} r_{jh} r_{kh} \quad (19)$$

And

$$r = \frac{r_{jk} + r_{jh}}{2}, \quad \text{df} = n-3 \quad (20)$$

Where r_{jh} and r_{jh} are the correlation for two variables and r_{kh} is the correlation between h and k, this test used with confidence when the sample size exceeds 20 [49] as what we have in the result.

3.2 Data Sets

3.2.1 Corpus

Our data sets are part of Semantic Evaluation (SemEval) that is a series of evaluations of computational semantic analysis systems. That was mined from several resources [50]. In each data set there are 750 sentence pairs. The headlines data set is mined from news sources by European Media Monitor RSS feeds. The inter-tagger correlation is equal to 79.4 %. In this data set the number of words is 11,228 words and the average of sentence length is 7.5 words and after removed the stop words the average sentence length is 5.5.

The image data set is image descriptions from the PASCAL VOC-2008 subsets [50]. In PASCAL VOC-2008 data set there are 1,000 images and it has been used by a number of image description systems. The image captions of the data set are released under a Creative Commons Attribution-ShareAlike license. The inter-tagger correlation is equal to 83.6%. In this data set the number of words is 13,689 words and the average of sentence length is 9.1 words and after remove the stop word the approximate sentence length is 5.

The OnWN data set is mined from the sense definitions from both WordNet and OntoNotes. The inter-tagger correlation equal to 67.2% the reason for the low correlation is that, the two sentences in a pair belong to two different senses. In this data set the number of words is 11,617 words and the average of sentence length is 7.7 words and after remove the stop word the sentence length is 3.7.

The MSR-Video data set is constructed from Microsoft Research Video Description Corpus collected during the summer of 2010. Workers on Mechanical Turk were

paid to watch a short video snippet and they summarized the action in a single sentence. The inter-tagger correlation is equal to 88%. In this data set the number of words is 9,945 words and the approximate average of sentence length is 6.6 words and after remove the stop word the approximate sentence length is 3.4.

Table 2 summarizes our corpora in terms of number of sentences and words.

	Headlines	images	OnWN	MSR-vid
NO of words with stop words	11228	13689	11617	9945
Average sentence length	7.5	9.1	7.7	6.6
No of words without stop words	8308	7464	5518	5065
Approximate sentence length without stop words	5.5	5	3.7	3.4
NO of sentences	1500	1500	1500	1500
NO of compared sentence pairs	750	750	750	750
Inter-tagger correlation percentage	79.4	83.6	67.2	88

Table 2 Number of Sentences, Words and Unique Words in the Corpora.

The biggest advantage of these data sets than the others is the gold standard it assembled using mechanical Turk, it contains a score between 0 and 5 for each pair of sentences.

Kappa index for a ranking is applied and followed the method measuring agreement over judgments of translation quality [52]. With the following interpretations:

- (5) The two sentences are completely equivalent, as they mean the same thing.
- (4) The two sentences are mostly equivalent, but some unimportant details differ.
- (3) The two sentences are roughly equivalent, but some important
- (2) The two sentences are not equivalent, but share some details.
- (1) The two sentences are not equivalent, but are on the same topic.
- (0) The two sentences are on different topics.

3.2.2. Document preprocessing

It is used to decrease inflectional word forms to the base common form, document preprocessing techniques are different from language to language because of the morphological and grammatical reasons, and then different normalization processes should be applied to a specific need.

3.2.2.1. Tokenization

Tokenization is a process to create tokens by breaking the text stream into text elements like symbols and words, then removing the unwanted characters like punctuation to assure that the two compared sentences are matched in type [53].

3.2.2.2. Normalization

Text normalization and lowercase folding is applied on our corpus by transforming text into a standard format. Normalizing the corpus to guarantee performing the highest similarity between the compared sentences is achieved.

3.2.2.3. Stop words

Stop words are the most common function words in the text, they are discarded in some applications because problems like phrase searching in texts that includes stop words occur (e.g. the, is, at, and etc.) but other specific applications avoid removing it [54]. In our corpus stop words are removed to focus on the significant semantic words in the context.

3.2.3. Vector space model

It is an algebra model for representing any object, in this thesis it is used to represent the text document. The standard vector space model is a widely used model especially in information retrieval [19], to disambiguation entities across documents. Using this model, each sentence is extracted and stored as a vector of terms after finishing the document preprocessing [19].

3.3. Computing Similarity

This section illustrates the sentence similarity methods used in the implementation.

3.3.1. Cosine similarity method

As described in section 2.3.2.1 it is the most popular measure to find the similarity between two vectors, if the following two sentences are compared (quoted from our data set OnWN - SemEval 2014), (A) “move in a group or flock” and (B) “move as a crowd or in a group”. These sentences are represented as vectors, section 2.3.2.1, as shown in table 3 below. The cosine similarity between these two sentences is calculated as 0.6667.

	Flock	crowd	Move	Group
A	1	0	1	1
B	0	1	1	1

Table 3 Cosine Similarity Calculation Example

3.3.2 Semantic similarity matrix method

As described before in section 2.3.2.5 semantic textual similarity matrix between two sentences, formally, according to the lexical similarity measure section 2.3.1 for each a_i and b_j word element w_{ij} . Word semantic similarity functions are used to find the similarity between pattern elements. Symmetric matrix founded when this measure is symmetric, ($w_{ij} = w_{ji}$), also self-similarity founded in the diagonal elements and it should have the greatest values.

Table 4 below shows the same sentences used in the example in section 3.3.1 to find the similarity between two sentences using the similarity matrix with words semantic similarity function section 2.3.1, the similarity between these two sentences is equal to 0.7669. (Self-similarity between words in sentences A and B is 1). If the similarity matrix was not used the similarity equal to the same as cosine similarity 0.6667.

W		S1	S2	S1,2	S1,2
		Flock	crowd	Move	Group
S1	flock	1	0.22	0	0
S2	crowd	0.22	1	0	0
S1,2	move	0	0	1	0
S1,2	group	0	0	0	1

Table 4 Word Matrix Score for the two Sentences, Using LSA Metric.

S1	1	0	1	1
S2	0	1	1	1

Table 5 Sentence 1 and 2 Vectors

Four word semantic similarity functions are used in the experiments, one of them is Corpus-based LSA as used in the example above and the remaining are WordNet-based LIN, WUP, and HSO.

3.3.3. Word order similarity method

Word order should be considered in similarity calculations. If two sentences are compared, S1 “The dog jumps over the fox.” and S2 “The fox jumps over the dog.” when the same similarity measurement in section 3.3.1 is applied, it shows that these two sentences are exactly the same as they contain exactly the same words. But in fact, these two sentences do not have the same meaning because the words dog and fox appear in different positions. A human can recognize the impact of order in meaning. Therefore, to be effective in computing the similarity between two sentences, word order must be accounted for.

When calculating word order similarity for the same example in section 3.3.1, the similarity is equal to 0.7226. In calculations, first, generate the joint word set for two sentences, table 6, then word order vectors are generated for each sentence depending on this joint word set indexing. The vector length is normalized by adding zero to the end, in case the word length is different between two sentences, then word

order formula is used to calculate the structure information between two vectors in table 6.

W	Move	Group	Flock	crowd
	1	2	3	4

Table 6 Joint Word Set.

S1	1	2	3
S2	1	4	2

Table 7 Word Order Example.

3.3.4. Overall sentence similarity method

Overall sentence similarity method is a combination of both word order similarity section 2.3.2.6 and cosine similarity section 2.3.2.1. with coefficient λ Section 2.3.2.7. The Overall similarity for the example explained in section 3.3.1 for all expected λ (Where λ is between 1 and 0.5) founded empirically. As shown in table 8 below:

1	0.9	0.8	0.7	0.6	0.5
0.7226	0.6723	0.6779	0.6835	0.6891	0.6946

Table 8 $S_{overall}$ Between two Sentences

3.3.5. Enhanced overall sentence similarity

A new method is proposed by enhancing the similarity formula explained in section 2.3.2.7, by changing the cosine similarity part, section 2.3.2.1, by the semantic similarity matrix part, section 2.3.2.5, as shown in the formula below.

$$SE_{overall}(S_1, S_2) = \lambda \frac{\vec{a} W \cdot \vec{b}^T}{|\vec{a}| |\vec{b}|} + (1 - \lambda) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (21)$$

$$SE_{overall}(S_1, S_2) = \lambda S_{se} + (1 - \lambda) S_r \quad (22)$$

Where SE represents the enhanced formula using the semantic similarity, S_{se} is used instead of using the cosine similarity.

For the example in section 3.3.1, the $SE_{overall}$ similarity is shown in table 9.

1	0.9	0.8	0.7	0.6	0.5
0.7669	0.7625	0.7581	0.7536	0.7492	0.74475

Table 9 $SE_{overall}$ Between two Sentences

When we compare both results using the two formulas $S_{overall}$ and $SE_{overall}$, the difference is clear as shown in table 10.

λ Value	$S_{overall}$	$SE_{overall}$
0.5	0.6946	0.7447
0.6	0.6891	0.7492
0.7	0.6835	0.7536
0.8	0.6779	0.7581
0.9	0.6723	0.7625
1	0.6667	0.7669

Table 10 $S_{overall}$ and $SE_{overall}$ Between two Sentences

In Table 11, some quoted sentences from our data set are used to show a small comparison between applying both overall similarity formula and the enhanced overall similarity, as described in sections 3.3.4, and section 3.3.5, using λ equal to (0.8).

Example	Enhanced Overall	Overall
A man is playing a large flute. A man is playing a flute.	0.84	0.80
The man hit the other man with a stick. The man spanked the other man with a stick.	0.76	0.68
A lion is playing with people. A lion is playing with two men.	0.75	0.70
A man and woman are driving down the street in a jeep. A man and woman are driving down the road in an open air vehicle.	0.82	0.48
A woman is applying eye liner to her eyelid using an eye pencil. A woman is applying cosmetics to her eyelid.	0.77	0.57

Table 11 Overall Similarity and Enhanced Overall Similarity Examples.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Experiments

This section analyzes the experimental results from the selected methods and data set, as explained in chapter three. Using four data sets collected from different fields gave a various range of writing expression, to challenge the selected formulas and find sentence similarity to compare.

4.2 Compare results

4.2.1 Comparison of lexical semantic relatedness with cosine similarity

As described, the two formulas in section 2.3.2.1 and section 2.3.2.5, also as an instance, section 3.3.1 and section 3.3.2, these formulas are applied on four different corpuses, section 3.2.1. Table 12 shows the obtained Pearson's correlations after comparing the method similarity with the human judge result.

Method Name	Headlines	Images	OnWN	MSRvid
Cosine Similarity	0.636	0.733	0.636	0.7141
Semantic Matrix – LSA	0.573	0.708	0.779	0.800
Semantic Matrix – LIN	0.550	0.463	0.600	0.603
Semantic Matrix – WUP	0.301	0.136	0.496	0.312
Semantic Matrix – HSO	0.155	0.131	0.306	0.142

Table 12 Pearson's Correlations for Cosine Similarity and Semantic Matric.

In Table 12 headlines and images corpus show the highest correlation in using cosine similarity than the semantic matrix methods for each corpus, even when corpus-based and WordNet-based word semantic similarity function were used. On the other hand, semantic matrix using corpus-based word semantic function LSA has the

highest Pearson's correlations for both headlines and images than the other semantic matrix methods which used WordNet-based word semantic similarity functions.

The OnWN and MSRvid corpus, Corpus-based word semantic similarity function LSA obtained the highest Pearson's correlation than cosine similarity and the semantic matrix using word semantic similarity function LIN, WUP and HSO.

When Pearson's correlations are calculated for all corpuses together in Table 13, Corpus-based word semantic similarity function LSA obtained the highest Pearson's correlations than the cosine similarity and other WordNet-based word semantic similarity function LIN, WUP and HSO.

Method Name	All Corpuses
Cosine Similarity	0.664
Semantic Matrix – LSA	0.719
Semantic Matrix - LIN	0.504
Semantic Matrix - WUP	0.287
Semantic Matrix - HSO	0.187

Table 13 Pearson's Correlations for all Corpuses for Cosine Similarity and Semantic Metrics

Based on the computed p-value for cosine similarity and semantic matrix using corpus-based word similarity function LSA equal to 0.000168, which is less than the significant level α equal to 0.05, this indicates that there is a significant difference between the cosine similarity method and the semantic matrix method using corpus-based word semantic similarity function LSA, which means using this method is better than the others.

Based on the results in table 13, it can be concluded that Corpus-based semantic matrix similarity using LSA is the best method and strongly related with human judge result than the cosine similarity method and the semantic matrix method using WordNet-based word semantic similarity function LIN, WUP and HSO.

The difference between the cosine similarity and the semantic matrix method is that the semantic matrix is added to the formula of the cosine similarity and this matrix improved the result significantly.

4.2.2 Overall similarity and enhanced overall similarity

As described, the formula in section 2.3.2.7 and as an instance in section 3.3.4 and section 3.3.5, four different data sets were applied, section 3.2.1, and Pearson's correlations results were obtained as in Table 14.

Method Name with ($\lambda = 0.8$)	Headlines	Images	OnWN	MSRvid
Overall Similarity	0.622	0.727	0.6359	0.707
Enhanced Overall – LSA	0.571	0.7085	0.766	0.786
Enhanced Overall - LIN	0.543	0.478	0.605	0.601
Enhanced Overall – WUP	0.313	0.154	0.504	0.329
Enhanced Overall – HSO	0.162	0.136	0.309	0.148

Table 14 Pearson's Correlations to Overall Similarity and Enhanced Overall Similarity.

Using overall similarity method, the headlines corpus obtained the highest Pearson's similarity than the other enhanced similarity methods. Using overall similarity method, the image corpus obtained the highest Pearson's correlations.

Comparing Pearson's correlations with the enhanced overall similarity method using word semantic similarity function LSA and calculating P-value equal to 0.068 and comparing with the significant level α equal to 0.05, lead to the conclusion that there is no significant differences with the enhanced overall similarity method using corpus-based word semantic similarity function LSA.

Both OnWN and MSRvid corpuses obtained the highest Pearson's correlations using the enhanced overall similarity method with corpus-based word semantic similarity function LSA. After comparing with the overall similarity and calculating p-value equals to 0.000001 for both corpuses. That means it is a significant difference than

overall similarity method and the other enhanced overall methods. (i.e. the linear relationship between the enhanced overall semantic similarity method using corpus-based LSA and the human judges result is strongly linear related than the relationship between the overall similarity and other enhanced overall similarity methods with the human judges results).

Also, if we calculate Pearson’s correlations for all corpuses together, as shown in table 15, the enhanced overall similarity method using corpus-based word semantic similarity function LSA, has the highest Pearson’s correlations than the overall similarity and the other enhanced semantic similarity methods. Then this means the enhanced semantic similarity method using word order semantic function LSA is strongly related with human than the overall similarity, and other enhanced overall semantic similarity methods.

Method Name with ($\lambda = 0.8$)	All Corpuses
Overall Similarity	0.661
Enhanced Overall Corpus-based metric – LSA	0.709
Enhanced Overall WordNet-based - LIN	0.508
Enhanced Overall WordNet-based – WUP	0.298
Enhanced Overall WordNet-based – HSO	0.192

Table 15 Pearson’s Correlations for all Corpuses Using Overall Similarity and Enhanced Overall Semantic Similarity.

4.2.3. All corpuses comparisons and ranking results

A Pearson correlation value comparison between human judge result and each formula is calculated, and then the results are sorted to rank and show the full map of our calculations.

For individual corpuses tables 16, 17, 18 and 19 show Pearson’s correlations. Table 20 shows all corpuses together.

	Headlines	Method Rank
Cosine Similarity	0.636	1
Overall Similarity	0.622	2
Semantic Matrix - LSA	0.573	3
Overall Similarity -LSA	0.571	4
Semantic Matrix - LIN	0.550	5
Overall Similarity -LIN	0.543	6
Overall Similarity -WUP	0.313	7
Semantic Matrix - WUP	0.301	8
Overall Similarity -HSO	0.162	9
Semantic Matrix - HSO	0.155	10

Table 16 Headlines Pearson Correlation Rank for all Methods

Table 16 headline corpus shows that using cosine similarity method has the highest Pearson correlation then using the overall similarity method. On the other hand, the semantic similarity matrix method using corpus-based word semantic similarity function LSA obtained the highest similarity than the enhanced overall similarity using the same word similarity function and the all other methods. Moreover, both semantic similarity matrix LSA and LIN obtained higher rank than the enhanced overall similarity methods using the same functions.

Table 17 images corpus also shows that cosine similarity method obtained the highest Pearson correlation than the other methods, but compared with the headline corpus ranked methods using the enhanced overall semantic similarity with LIN function obtained higher Pearson correlation than the semantic matrix methods using the same function.

	Images	Method Rank
Cosine Similarity	0.733	1
Overall Similarity	0.727	2
Semantic Matrix - LSA	0.708	3
Overall Similarity -LSA	0.708	4
Overall Similarity -LIN	0.478	5
Semantic Matrix - LIN	0.463	6
Overall Similarity -WUP	0.154	7
Semantic Matrix - WUP	0.136	8
Overall Similarity -HSO	0.136	9
Semantic Matrix - HSO	0.131	10

Table 17 Images Pearson Correlation Rank for all Methods.

Table 18 OnWN corpus shows semantic matrix method using LSA word semantic similarity function obtained the highest Pearson correlation than all other methods. The enhanced semantic similarity with the same function obtained higher correlation than the other methods except the semantic matrix with the same function (i.e. LSA).

	OnWN	Method Rank
Semantic Matrix - LSA	0.779	1
Overall Similarity - LSA	0.766	2
Cosine Similarity	0.636	3
Overall Similarity	0.636	4
Semantic Matrix - LIN	0.600	5
Overall Similarity - LIN	0.504	6
Overall Similarity - WUP	0.504	7
Semantic Matrix - WUP	0.496	8
Overall Similarity - HSO	0.309	9
Semantic Matrix - HSO	0.306	10

Table 18 OnWN Pearson Correlation Rank for all Methods.

	MSRvid	Method Rank
Semantic Matrix - LSA	0.800	1
Overall Similarity - LSA	0.786	2
Cosine Simialrity	0.714	3
Overall Similarity	0.707	4
Semantic Matrix - LIN	0.603	5
Overall Similarity - LIN	0.601	6
Semantic Matrix - WUP	0.312	7
Overall Similarity - WUP	0.329	8
Semantic Matrix - HSO	0.142	9
Overall Similarity - HSO	0.148	10

Table 19 MSRvid Pearson Correlation Rank for all Methods.

Table 19 MSRvid corpus, semantic matrix using LSA word semantic similarity function obtained the highest Pearson correlation than all other methods. Compared with the other corpuses, semantic matrix using both WUP and HSO function obtained higher Pearson correlation than the enhanced overall semantic similarity using the same word semantic similarity function.

On the other hand, comparing semantic matrix methods with the enhanced overall semantic methods, both use the same word semantic similarity function. We can say using semantic matrix with LSA function obtained higher similarity than the enhanced semantic method using the same semantic function. And for LIN function, semantic matrix obtained higher Pearson correlation than the enhanced method using the same function, except in image corpus, the case is opposite. In both WUP and HSO functions, using semantic matrix obtained the higher similarity than the enhanced overall similarity methods using the same semantic function except in MSRvid corpus. Moreover, cosine similarity compared with the overall similarity method has higher Pearson correlation.

	All corpuses	Method Rank
Semantic Matrix - LSA	0.719	1
Overall Similarity - LSA	0.709	2
Cosine Similarity	0.664	3
Overall Similarity	0.661	4
Overall Similarity - LIN	0.508	5
Semantic Matrix - LIN	0.504	6
Overall Similarity - WUP	0.298	7
Semantic Matrix - WUP	0.287	8
Overall Similarity - HSO	0.192	9
Semantic Matrix - HSO	0.187	10

Table 20 Pearson Correlation Rank for all Corpuses.

Table 20 compares and ranks all corpuses together. The results show using semantic matrix with word semantic similarity function LSA obtained the highest Pearson correlation than all other methods. On the other hand, all the enhanced semantic similarity methods using WordNet-based word semantic similarity functions obtained higher Pearson correlation than the semantic matrix method using the same function.

4.2.4 Compare with other methods using the same data sets

To compare this work with other researchers who used the same data sets but different methods, Table 21 below summarizes our comparisons.

Metrics names		Headlines	Images	OnWN	MSRvid
Other methods range	Maximum	0.7837	0.8214	0.8745	0.8803
	Minimum	0.0177	0.3243	0.3607	0.0057
Cosine Similarity		0.636	0.733	0.6364	0.714
Corpus-Based Word function and Semantic Matric	LSA	0.573	0.708	0.779	0.800
WordNet-Based Word function and Semantic Matric	LIN	0.550	0.463	0.600	0.603
	WUP	0.301	0.136	0.496	0.312
	HSA	0.155	0.131	0.306	0.142
Overall Similarity		0.622	0.727	0.6359	0.707
Enhances-Overall Similarity (LSA)		0.571	0.708	0.766	0.786
Enhances-Overall Similarity (LIN)		0.543	0.478	0.504	0.601
Enhances-Overall Similarity (WUP)		0.313	0.154	0.504	0.329
Enhances-Overall Similarity (HSA)		0.162	0.136	0.309	0.148

Table 21 Compare With Other Results that use the Same Data Sets.

Different results were obtained by other researchers who used different techniques to find the sentence similarity and consider additional factors like using grammatical functions, etc. in the calculations [46].

Table 22 shows comparisons and ranking of the thesis results with other researchers.

Corpus Name	Maximum	Minimum	System best value	System rank with others
Headlines	0.7837	0.018	0.636	21 from 38
Images	0.8214	0.324	0.733	13 from 38
OnWN	0.8745	0.361	0.779	18 from 38
MSRvid	0.8803	0.006	0.800	24 from 90

Table 22 Results Rank with other Researchers

For headlines data sets from 2014, mined from news sources by European Media Monitor using the RSS feed, the thesis results ranked in the third quarter. For image data sets from 2014, which is image description from the PASCAL VOC-2008 subsets, the thesis results ranked in the second quarter. For OnWN from 2014, that is mined from the sense definitions from both WordNet and OntoNotes, the thesis results ranked in the second quarter. Finally for the MSR-Video from 2012, it is a Microsoft Research Video Description Corpus collected from 2010, the thesis results ranked in the first quarter compared with other methods using the same data sets.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1. Conclusion

This work has successfully addressed all the aims and objectives of the research. Different methods that have been developed to calculate the similarity between two sentences are selected and compared with each other using different data sets.

The results show using semantic matrix method with corpus-based word semantic similarity function LSA, achieved a significant performance compared to using the same method with WordNet-based word semantic similarity functions (i.e. LIN, WUP and HSO). This means, adding the word semantic function to the cosine similarity method achieve a significant performance. On the other hand, adding the word order similarity to the semantic function method achieved a significant difference, except when WordNet-based word semantic similarity function was used.

The results show adding word semantic similarity function into the introduced overall similarity method achieved a significant performance compared with the overall similarity method that does not use the word semantic similarity function. These methods with different word semantic relatedness function have been compared. Also, the results show using Corpus-Based word semantic relatedness function significantly improved the similarity result compared to using WordNet-based word semantic relatedness function.

The results of this study light the way and give a comparison between these methods using the same data sets that are selected from different fields to test these methods into a wide range of language expressions.

5.2. Future works

- Applying the same formulas on different languages after preparing other language required materials for implementations.
- Deeply analyzing the differences between the selected corpuses to find why the results are different.
- Analyzing the compared sentences before selecting the suitable similarity formula.
- Applying the hybrid word semantic similarity function on the same data sets then comparing the results with the used WordNet and corpus based word semantic similarity functions.

REFERENCES

1. **Cilibrasi R., Vitányi P., (2006)**, “*The Google Similarity Distance*”, IEEE Trans Know Data Engineering, pp. 370-383.
2. **Batet M., (2011)**, “*Ontology-Based Semantic Clustering*”, AI Communication 24, pp. 291-292.
3. **Jones K., Walker S. and Robertson S., (2000)**, “*A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. Part*”, in Information Processing and Management, pp. 779–808.
4. **Barzilay R. and Elhadad M., (1997)**, “*Using Lexical Chains for Text Summarization*”, in Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 10-17.
5. **Mehran S. and Timothy H., (2006)**, “*A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets*”, In WWW 06 ACM Press, pp. 18.
6. **Yuhua L., Zuhair B., David M. and James O., (2006)**, “*A Method for Measuring Sentence Similarity and its Application to Conversational Agents*”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 8.
7. **Murphy M., (2003)**, “*Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*”, Cambridge: Cambridge University Press, pp. 22,137.
8. **Cruse D., (1986)**, “*Nottingham Linguistic Circular*”, On Lexical Ambiguity, Lexical Semantics, Cambridge: Cambridge University Press, pp. 65–80.

9. **Hirst G. (1995)**, “*In Working Notes, AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generality*”, Near-Synonymy and the Structure of Lexical Knowledge, Stanford, pp. 51-56.
10. **Scott M., (1999)**, “*WordSmith Tools*”, Version 3, Oxford: Oxford University Press, pp. 65-103.
11. **Marti A. and Schütze H., (1993)**, “*Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*”, Customizing a Lexicon to Better Suit a Computational Task, In Columbus, OH USA, pp. 55–69.
12. **Kirkpatrick B., (1998)**, “*Roget’s Thesaurus of English Words and Phrases*”, Harmondsworth, Middlesex, England: Penguin, pp. 36.
13. **NLM. U.S. National Library of Medicine. (2004)**, “*Medical Subject Headings*”, pp. 54-72.
14. **Spitzer M. and Lowe C., (1998)**, “*Information Literacy: Essentials Skills for the Information Age*”, Information Resources Publications, Syracuse University, Syracuse, NY, USA, pp. 18.
15. **LCSH U.S. Library of Congress, 26th Edition, (2003)**, “*Library of Congress Subject Headings*”.
16. **Christiane F., (1998)**, “*WordNet: An Electronic Lexical Database*”. MIT Press, pp. 147–165.
17. **Raghavan P. and Schütze H., (2008)**, “*Introduction to Information Retrieval*”, Cambridge University Press, pp.13.
18. **Gerald S., (1989)**, “*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*”, Addison-Wesley, Reading, pp. 18.

19. **Baeza-Yates R., Ribeiro-Neto B. (1999)**, “*Modern Information Retrieval*” Harlow: ACM Press/Addison-Wesley, pp. 54-72.
20. **Li X. and Croft W., (2003)**, “*Time-Based Language Models*”, In Proceedings of CIKM, pp. 32.
21. **Dipanjan D., Desai C., André F., Nathan S., and Noah A., (2014)**, “*Frame-Semantic Parsing. Computational Linguistics*”, Association for Computational Linguistics, pp. 40.
22. **Jeffrey F., Sam T., Jaime C., Chris D., and Noah A., (2014)**, “*A Discriminative Graph-Based Parser for the Abstract Meaning Representation*”, Baltimore, MD, USA, pp. 54-72.
23. **Manfred S., (2011)**, “*Discourse Processing*”, Number 15 in Synthesis Lectures on Human Language Technologies, Morgan & Claypool, San Rafael, CA, pp. 36.
24. **Claudia L. and Martin C., (1998)**, “*Combining Local Context and WordNet Similarity for Word Sense Identification*”, In Christine Fellbaum Editor, WordNet – An Electronic Lexical Database, MIT Press, pp. 18.
25. **Vassiliki J. K. and Achilleas V., (2007)**, “*Fuzzy Reasoning in the Development of Geographical Information Systems*”, International Journal of Geographical Information Systems, pp. 209-223.
26. **Hirschman L., (2001)**, “*Natural Language Question Answering the View from Here*”, Natural Language Engineering, Cambridge University Press DOI, United Kingdom, pp. 275–300.
27. **Leacock C., Chodorow M. and Miller G., (1998)**, “*Using Semantics and WordNet Relation for Sense Identification*”, Association for Computational Linguistics, pp.18.

28. **Wu Z. and Palmer M., (1994)**, “*Verb semantics and Lexical Selection*”, In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 133-138.
29. **Resnik P., (1995)**, “*Using Information Content to Evaluate Semantic Similarity in a Taxonomy*”, In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453.
30. **Francis W. and Henry K., (1982)**, “*Frequency Analysis of English Usage*”, Lexicon and Grammar, Houghton Mifflin, Boston, pp. 72.
31. **Lin, D., (1998)**, “*An Information-Theoretic Definition of Similarity*”, In Proceedings of the International Conference on Machine Learning, pp. 18.
32. **Jay J. and David W., (1997)**, “*Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*” In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), Taiwan, pp. 54.
33. **Choueka Y. and Lusignan S., (1985)**, “*Disambiguation by Short Contexts Computers and the Humanities*”, pp. 147–157.
34. **Satanjeev B. and Ted P., (2003)**, “*Extended Gloss Overlaps as a Measure of Semantic Relatedness*”, In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 805–810.
35. **Hirst G. and St-Onge D., (1995)**, “*Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*”, Edited by Christiane Fellbaum, Cambridge, MA: The MIT Press, pp. 72.
36. **Thomas K. and Susan T., (1997)**, “*A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge*”, American Psychological Association, vol. 1, pp. 211-240.
37. **Turney P., (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. 324.

38. **Chklovski T. and Pantel P., (2004)**, “*Verbocean: Mining the Web for fine-Grained Semantic Verb Relations*”, In Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 18.
39. **Rapp R., (2004)**, “*Discovering the Senses of an Ambiguous Word by Clustering its Local Contexts*,”, in Proceedings of the 28th Annual Conference of the Gesellschaft fur Klassifikation, pp. 521-528.
40. **Landauer T. and Dumais S., (1997)**, “*A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge*”, Psychological Review, vol. 104, pp. 211-240.
41. **Zaka B., (2009)**, “*Theory and Applications of Similarity Detection Techniques*”, Institute for Information Systems and Computer Media (IICM) Graz University of Technology A-8010 Graz, Austria, pp. 342.
42. **Long Q., Min-Yen K. and Tat-Seng C., (2006)**, “*Paraphrase Recognition via Dissimilarity Significance Classification*”, Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 18–26.
43. **Yitao Z. and Jon P., (2005)**, “*Paraphrase Identification by Text Canonicalization*”, In Proceedings of the Australasian Language Technology Workshop 2005, pp. 160–166, Sydney, Australia, pp. 72.
44. **Rada M., Courtney C. and Carlo S., (2006)**, “*Corpus-based and Knowledge-based Measures of Text Semantic Similarity*”, American Association for Artificial Intelligence, pp. 36.
45. **Spärck K. J. (1972)**, “*A Statistical Interpretation of Term Specificity and its Application in Retrieval*”, Journal of documentation, pp. 11–21.
46. **Samuel F. and Stevenson M., (2007)**, “*A Semantic Similarity Approach to Paraphrase Detection*”, pp. 18-54.
47. **Yuhua L., Zuhair B., David M. and James O., (2006)**, “*A Method for Measuring Sentence Similarity and its Application to Conversational*

Agents”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 8.

48. **Li J., Bandar Z., McLean D. and Shea O., (2004)**, “*A Method for Measuring Sentence Similarity and its Application to Conversational Agents*”, 17th International Florida Artificial Intelligence Research Society Conference, Miami Beach, AAAI Press, pp. 820–825.
49. **Lawrence I. and Lin K. (1989)**, “*A Concordance Correlation Coefficient to Evaluate Reproducibility*”, Biometrics, pp. 255–268.
50. **SemEval-Semantic Evaluation, (2014)**, “*Proceedings of the Workshop*” The 8th International Workshop on Semantic Evaluation, pp. 72-90.
51. **Richard S., Andrej K., Quoc V. L. and Christopher D., (2014)**, “*Manning, and Andrew Y. Ng. 2014, Grounded Compositional Semantics for Finding and Describing Images with Sentences*”, Transactions of the Association for Computational Linguistics, pp. 207–218.
52. **Ergun B. and Andy W., (2014)**, “*Referential Translation Machines for Predicting Translation Quality*”, In Ninth Workshop on Statistical Machine Translation, Baltimore, USA, Association for Computational Linguistics, pp. 54.
53. **Grefenstette G. and Segond F., (1997)**, “*Multilingual Natural Language Processing*”, International Corpus of Corpus Linguistics, pp. 2.
54. **Ellen R., (2011)**, “*Little Words can Make a Big Difference for Text Classification*”, International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 130-136.

APPENDICES A

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: KAIRALDEEN, Ammar Riadh

Date and Place of Birth: 21 November 1981

Marital Status: Married

Phone: +90 534 410 84 56

Email: eng_ammarr81@yahoo.com



EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	Çankaya University in Computer Engineering Department.	December 2014
Higher Diploma	Iraqi Commission for Computers and Informatics in Information Technology / Website Technique Department.	November 2010
B.Sc.	Electric and Electronic Technical College in Computer Engineering Department.	September 2005

FOREIN LANGUAGES

Arabic, English