Journal of Computer Science

Original Research Paper

New Information Content Glossary Relatedness (ICGR) Approach for Short Text Similarity (STS) Tasks

Ali Muftah BenOmran and Mohd Juzaiddin Ab Aziz

Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, 43000, Selangore, Malaysia

Article history
Received: 10-03-2018
Revised: 14-05-2018
Accepted: 04-6-2018

Corresponding Author:
Ali Muftah Benomran
Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, 43000, Selangore, Malaysia
Email: aliomran@siswa.ukm.edu.my

Abstract: The measurement of the relatedness of word semantics based on complementary Wikipedia and WordNet-based methods takes two forms, combined and integrative, which are aimed at increasing the semantic space between related words. However, each form has its own set of issues regarding its components and the strategy that is used to combine and integrate corpus-based and knowledge-based methods. In the integrative strategy, a large corpus, such as Wikipedia, is used to extract a set of related words for a particular concept as a basis for searching the WordNet space. The drawback to this strategy is in its use of a fixed scaling parameter, which only fits an implemented dataset that is near to a human score. Other corpus-based methods use a cut-off threshold that is determined experimentally to reduce the semantic space and to increase the search for a more accurate semantic space. Such methods merely take into account the frequency of bigrams, while ignoring the frequency of individual terms. Knowledge-based methods using a gloss overlap have a similar limitation to the corpus-based methods, where they lead to the loss of many valuable relatedness features that determine a more accurate measurement. Thus, in this paper, a new Information Content Glossary Relatedness (ICGR) approach was proposed in two steps, namely, an Extended-PMI based on a cut-off density threshold was proposed to extract a Robust Relatedness Vector set (RVS) of a large Wikipedia dataset. Then, a Semantic Structural Information (SSI) method was presented to use the RVS as a fulcrum to define the most relatedness gloss in the WordNet of each gloss and to select the top 5 glosses related to each RVS. The results showed that the proposed approach outperformed the state-of-the-art set, where the Extended-PMI achieved a Spearman's correlation of 0.89 to the human score and the ICGR approach achieved a Spearman's correlation of 0.8 to the human score.

Keywords: Wikipedia, PMI, WordNet, Gloss, Structural Information, STS

Introduction

The similarity of words expressed in a natural language is a challenge in NLP in several domains. It is often decompressed into a comparison of the semantic relations between concepts, depending on knowledge-based corpora such as Wikipedia and WordNet. Knowledge-based methods are bound by the content of the terminological resource, while the context might contain additional content not covered by them (Jimeno-Yepes and Aronson, 2012). Such methods derive word-to-word semantic relatedness by searching the taxonomy of four features, namely, path and depth-based, content-based, gloss-based and combined approaches (Aouicha *et al*., 2016), where similarities can be detected through several semantic features such as those thatare measured by detecting elements of homonyms, synonyms, meronyms, holonyms, hypernyms, hyponyms and antonyms(Alzahrani, 2016; De Luca and Nürnberger, 2006).

A gloss-based approach has been shown to outperform most of the existing approaches as it considers almost all the features in WordNet (Zesch *et al*., 2016). All gloss-based methods based on WordNet stem from (Lesk, 1986), which inspired a whole family of methods that exploited the number of common words in two sense definitions (glosses) to select the proper meaning in a context. Glosses play a key role in the Lesk

algorithm, which exploits only two types of information: (1) A set of dictionary entries, one for each possible word meaning and(2) information about the context in which the word occurs. This approach suffers from two problems: (1) complexity - the number of comparisons increases in combination with the number of words in a text; and (2) definition expressiveness - the overlap is based only on word co-occurrences in glosses. The first problem is tackled by a "simplified" version of the Lesk algorithm (Kilgarriff and Rosenzweig, 2000), which disambiguates each word separately. When given a word, the meaning of the gloss that shows the maximum overlap with the current context, as represented by the surrounding words, is selected. The simplified Lesk algorithm significantly outperforms the original Lesk algorithm, as was proven by (Kilgarriff and Rosenzweig, 2000). The second problem was faced by (Banerjee and Pedersen, 2003), who proposed an "adapted" Lesk algorithm. The adapted version exploits the relationships among meanings: Each gloss is extended by the definitions of semantically-related meanings. these issues discussed in many domains close relate to STS approaches such asQuestion Answering System and Semantic conflicts detectionof limited context (Al-Baltah *et al*., 2014; Sweta and Singh, 2017)

De Luca and Nürnberger (2006) talk about the sense disambiguation problem using lexical resources to perceive the meaning of a query word. This meaning could then be used to retrieve the related relevant documents based on hyponyms, as well as further context information like domain and glosses. The retrieved document will then undergo processing and representation as vectors in an m-dimensional space through the vector space model. Then, the documents are classified by calculating their similarity to every prototype vector that is used to give descriptions to the Sense Folders. Here, every word in the semantic space has a relationship with the words of noun collocations or what is referred to as the WordNet ontology. The k-means clustering algorithm was then used to search a subset of documents for every meaning possible within each prototype vector. The combined approach of WordNet and Sense Folders performed better than the individual implementation. However, the SynSets that was categorised as the "Factotum" domain had a strong influence on the classification performance. The "rule" example illustrated that in the absence of any merging methods, the categorisation performance is poor because the word sense distinction was too finely grained and is unusable for information retrieval purposes. Furthermore, the SynSets labelled as 'Factotum' entry were too general to be merged and were not very expressive.

On the other hand, Rodríguez and Egenhofer (2003) calculated the similarity of terms using the weighted sum of the similarity of their corresponding concepts' three features: The conceptual features (e.g., attributes, meronyms, etc.), degree of overlap between synonym setsand neighbour concepts (i.e., those who possess a path distance to the evaluated concept that is equivalent to or lower than a natural number r).The feature-based similarity measure that (Tversky, 1977) defined served as the basis for the calculation of each of these three components. This measure accounts for the relative importance between non-common and common characteristics.

Petrakis and Varelas (2006) created an extension of the previous approach by introducing X-similarity, which depended on the matching that is found between synonym sets and concept glosses that have been taken from WordNet (i.e., words that were taken through parsing concept definitions) or scope notes that were taken from MeSH. Once each comparison term is individually matched to each ontology's concept, they are taken to be similar if their synonyms (i.e., different labels that represent a single ontological concept), glosses and the concepts within their vicinity have lexical similarity. Calculation of the similarity is done by using the maximum similarity value that had been obtained by making a comparison of the synonyms and glosses per separate through the Jaccard coefficient (Jaccard, 1982). Nevertheless, they are typically dependent on non-taxonomic features that are normally not found in ontologies (Chen *et al*., 2009). Furthermore, they require the weighting parameters to be fine-tuned for the heterogeneous semantic evidences to be integrated (Petrakis and Varelas, 2006). Generally, these approaches suffer from the Lowest Common Sub-sumer (LCS), where any two words that possess the same depth will be considered to have very high similarity. Moreover, the adjusted parameters are utilised for the integration of heterogeneous semantic evidences.

Aouicha and Taieb (2015) introduced new G2WS approaches. In their methods, Wiktionary was used to connect the gloss to their factors. Furthermore, noun factors were also calculated through the utilisation of the glosses of the ancestors that were allocated to a specific concept within the "is a" taxonomy of WordNet. An Information Content (IC) measure is used to obtain the factor of each noun. For each entry, the semantics is computed through the use of factors assigned to the nouns and referring to the set of glosses that have been taken from the subgraph of the ancestors (view Fig. 1) and supplemented by information taken from Wiktionary. In this context, WordNet and Wiktionary are used to supply glosses of the concepts that were utilised. Lastly, Noun Factor was calculated by exploiting function so that the factor allotted to a computed concept factor can be quantified and used to provide the degree of semantic similarity between two concepts. Spearman evaluation and Pearson correlation measurements were utilised in the evaluation of the semantic similarity using the following six benchmarks as the bases - RG65, AG203, MC30, MED38, GeReSiD50 and SimLex666. Compared to the state of the art, this method demonstrated very good correlation.

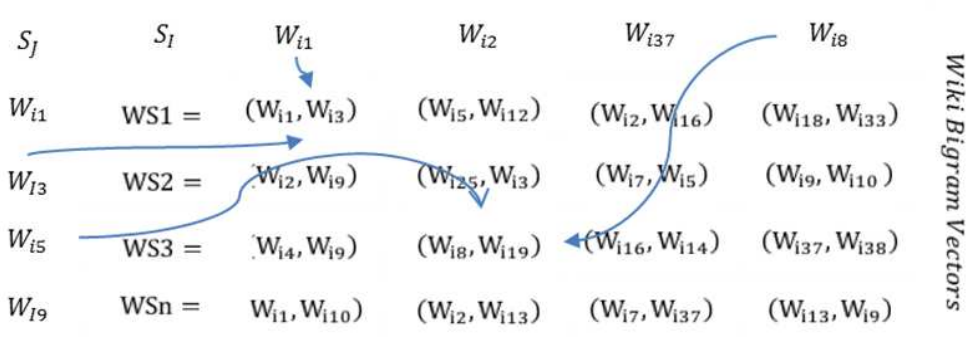


Fig. 1: Extract of wiki semantic set using the built bigram from the Wikipedia corpus, where W_I and W_j have sets in S_I and S_J with a list of related words

Aouicha *et al*. (2016) suggested the utilisation of gloss-based semantic similarity (G3W) using the Ancestors' subgraph-based IC quantification (AIC) as the basis. The estimation of AIC estimation for every ancestor contribution has its basis in the quantification of two parameters, a concept's depth, which pertains to specificity or generality and its hyponyms. However, the method on this section uses a general assumption that is based on the notion that being deeper within the ontological hierarchy makes something semantically richer compared to something that just stays on the top. For the "is a" hierarchical structure, the basic features from the superordinate concept are inherited by a subordinate concept before it adds its own specific features so that it could generate a new meaning. Lastly, the semantic similarity is computed using a set of terms found within the similarity space. The semantic relatedness approach is also utilised, which works based on the noun weighting technique using the IC-computing method. Furthermore, Wiktionary, WordNet and Wikipedia resources were utilised. The semantic similarity between the two concepts is first calculated. The overlap is used as a representation of the sum of the weights of the two concepts' overlapping nouns. 3GW is based on a weighting mechanism that is implemented on the nouns that have been taken from the glosses that were allocated to a target concept. The AIC computing method is used to compute for the noun of the weight. Then, the Dice coefficient was used to generate the similarity estimation through the use of the weights of over-lapped words. Experiments that were conducted on a set of datasets, including multiword expressions and verbs and pairs of nouns, showed that this measure performs better than all the quoted gloss-based measures.

Materials

A large corpus gives researchers a wider scope to infer and conclude the meanings of words from the repeated semantic patterns in a list of words. However, words can have more than one sense, with an unclear semantic distance between their actual meanings in large corpus. This affects the validity of the well-known theory of Firth (1957), who stated that words that occur in similar contexts tend to be close semantically (Hirst, 2006). Even though there are many techniques that can address this problem, they can hardly decode the implied semantics information of the context from the corpus into vector spaces (BenOmran and Aziz, 2014; Lucie, 2017). The disadvantage of this category is data sparseness, where the text is affected by its sparseness. To present the high semantic space of a large corpus, word specifications are derived from their co-occurrence distribution strategy (Bollegala *et al*., 2017). These co-occurring features of words are represented in a vector space for future computations of relatedness. The richness of document corpora is used for information extraction (e.g., bigrams), feeding relatedness functions and enhancing the accuracy of the output (Pesaranghader *et al*., 2015). Nevertheless, the imprecise and noisy nature of the extracted information demands a purification phase to remove those features that are insignificant for the calculation of more reliable estimations. However, due to the high overlap of previous methods, bigrams, as they appear together, need not necessarily be semantically related. The connection point between the semantic corpus, semantic space and the ontological glosses has not been well-presented (Jager *et al*., 2016). Instead, a simple bigram space, supported by cut-off thresholds to reduce the size of the vector, is presented. The second issue to note is that most of the gloss-based methods that have been presented ignore the structural relations of the word definitions inside the taxonomy and instead relate to the extracted bigram vectors of a large corpus to determine, in the presence of a whole definite set, the nearest gloss to the bigram vector (Jabeen, 2014).

Wikipedia Statistic Calculation Extraction

The quality of Wikipedia vectors depends on refining the strategy of the Wikipedia dataset, which is a necessary step to build a more representative corpus with

the required reliable features (Javanmardi and Lopes, 2010). Thus, pre-processing steps must be implemented: First, the Wikipedia articles must be processed into sentences, where the lengths of the sentences range from three words, for short texts, to a maximum of 25 words, for long texts.Each sentence is pre-processed by eliminating the stop words and keeping their order in the sentences to generate the bigrams and unigrams. To present the Wiki unigram, the frequency of all the terms in each sentence in the Wikipedia dataset is found using the following standard unigram formula in Equation 1:

$$Wiki\ Unigram\ Unif^{c}(w_i) = \frac{C(W_I)}{N} \quad (1)$$

The Wiki bigrams are calculated for each sentence in the processed data, where$f^{c}(w_i)$ represents the frequency in terms of each word in the Wikipedia taken after the pre-processing step. Here, $f^{c}(w_i)$ is defined in the entire corpus, $C(W_I)$ and divided by (N), which is the number of tokens, as follows:

$$(W_{IB})_{Wiki-bigram} = \begin{cases} for \forall WS \\ F(W_I|W_J) \in WS \\ sentence\ lenght W_i \geq SL \geq W_J \end{cases} \quad (2)$$

In this step, for each sentence$\forall(WS)$ in the Wikipedia statistics corpus, the bigrams are generated as$F(W_i|W_i$-1), where $W_i$is the last word in the sequence I andW_J is the previous word in the same sequence for the current sentence,WS_N. Each sentence is presented as a set of pairs of words that appears tighter.

Proposed ICGR Approach

Word2wiki-Vector (Wiki-2WV) Representation

Any word is part of a sentence, as the main aim is to measure the STS task with regard to the relatedness of the word semantics. In this step, the Word2wiki-vector (Wiki-2WV) is built from two short sentences, S_I and S_J, which have been pre-processed by removing the stop words and presenting the sentences as a Bag Of Words (BOW), S = $\{w_1,w_2,\ldots,w_n\}$, in which a sentence is considered as a common bag of words instead of a sequence of connected words. For each word in S_I and S_J, the unigrams and all possible bigrams from each sentence are selected. The bigram is a list of all the words that appear with the words in the processed S_I and S_J, as in the following Fig. 1.

In this step, the corpus statistics are generated once to any bigram element (w_{in}, w_{jm}) in the bigram list. The corpus statistics are used to define the bigram frequency f^{s} ($W_i|W_{BI}$) by computing the number of timesw_i and w_{Bi} appear in the same sentence. In addition, the corpus frequency, $f^{c}(w_i)$, of each bigram in the unigram list is computed using Equation 3 to present all the bigrams in each sentence as the Wikipedia Words to the Vectors (Wiki-2WV) and to determine their appearance in the corpus:

$$Wiki-2WV(S_I) = \left(SW_I', (W_{BI})_{Wiki-bigram}, N\right) \quad (3)$$

The Wiki-vectors of each word in S_I and S_J are presented by searching the Wiki-bigram space $(W_{BI})_{Wiki\text{-}bigram}$ to find the words that appear with W_S and their frequency of the appearance, N. For example, in Fig. 1, S_I has a set of six words. In this case, the word W_{i1} has two words, W_{i3} and W_{i10}, that appear once each in the Wiki-bigram.

Extended PMI based on Cut-Off Density Threshold

In this step, the standard PMI (Islam adn Inkpen, 2008) is extended with a cut-off density threshold. The Extended-PMI used to measure the Relatedness Vector Set (RVS) based on a two-part process. First, the threshold that has been trained on words from a short text is implemented to define the cut-off density threshold so as to decrease the unrelated RVS space in the generated corpus by training the Extended-PMI on a part of the dataset. For example, the SI wiki-vector is presented with an Extended-PMI to define the threshold density. Each word in the short text is defined by a collection of relatedness vector values, as in Fig. 1. Based on the cut-off threshold, the RVS accepts or rejects the words that appear with the relatedness vector of each word in the short text. The formula in Equation 4 is presented as follows:

$$RVS(SW_I) = \{SW_I : Wiki-2WV(S_I) \geq T\} \quad (4)$$

The Relatedness Vector Set (RVS) of (SW_I), the sentence S_I with word (W_I), is calculated using the Extended-PMI with the number of co-occurrences and individual occurrences of words (marginal frequencies) in the short text. The cut-off density threshold, Tr, is a factor for filtering out the words that have a low Extended-PMI relationship. The following Equation 5 presents the PMI under two conditions:

$$Extended-PMI\left(S_I(W_i), W_{Bi}\right) = \begin{cases} Wj \in Wiki\ Bigram(W_{Bi}) \\ \dfrac{\frac{1}{2} * \left(f^{s}\left(S_I(W_i), W_{Bi}\right) + f^{s}F\left(W_{Bi}S_I(W_i)\right)\right)}{\min\left(f^{c}\left(S_I(W_i)\right), f^{c}(W_{Bi})\right)} \\ * \dfrac{\left(f^{c}\left(S_I(W_i)\right) * f^{c}(W_{Bi})\right)}{maxf} \end{cases}$$

First, the Extended-PMI is calculated, where S_I= $W_{i1}W_{i3}W_{i7}W_{i37}W_{i16},W_{i33}$. The frequency, $f^e(S_I(W_i), W_{Bi})$ with Wiki-2WV(S_I) is calculated for each Wiki-word (W_i) in S_I to indicate the number of sentences. Then, ($f^e(S_I(W_i), W_{Bi})$ and $f^eF(W_{Bi}S_I(W_i))$) are calculated to cover all the possible appearances of W_i to the left or right side of the bigrams and are normalized with $f^e(S_I(W_i))f^e(W_{Bi})$ as the corpus frequency of the word W_{Bi} and wordW_i, respectively.

The process breaks down the traditional high-relation PMI towards infrequent words (rare words), which goes against the hypothesis of the semantic distribution similarity by using the minimum frequency between the sentence words W_i and the Wiki-2WV(S_I). This conservativeness is a limitation of the PMI. The Extended-PMI is normalized with the total frequency between the RVS in the corpus ($f^e(S_I(W_i))*f^e(W_{Bi})$) divided by the total frequency, (max$f$). This is aimed at giving a fixed upper bound score in case of a full match, which is equal to 1. In other words, SR-PMI is forced to take the maximum of 1 in the case of a positive dependence, i.e., when one word only occurs in the context of another, but not necessarily the other way around. A large-sized RVS is generated for each word in $S_I(W_i)$. Much of the RVS is generated with a low semantic PMI value, which is far from the high and related space of the RVS. The cut-off threshold (Tr) is determined as in the following equation:

$$Tr(S_I) = \frac{\forall\left(Extended - PMI\left(S_I(W_i)RVS\right)\right.}{NW_i\left(RVS\right)} \quad (6)$$

where, the Extended-PMI calculates the relatedness between the words (W_i) in the sentences and the bigram words in the Wikipedia corpus statistics. The RVS set is generated with a low semantic value, which is far from the high and related space. The $Tr(S_I)$ eliminatesvalues lower than the density threshold start to be sparse and far from the main word in the short text. Considering the following example,S_I = “Remains are floating on the shore after the sinking.” and S_J= “Bodies lay on the coast after the boat is drowned in the sea”. The Extended-PMI is implemented as follows in Tables 1 and 2.

According to the cut-off density threshold, the RVS words are: For S_I, Remains (dead, body, Human), Floating (buoyed, water, Liquid), Shore (fringe, Edge, large, body, Water) and Sinking (slow, fall, Decline); and for S_J, Bodies (physical, structure, animal, human, related, adjectives, corporeal, Physical), Lay (Put, down), Coast (area, land, Sea), Boat (small, vessel, travel, Water), Drown (Die, submersion, inhalation, water, liquid), Sea (division, ocean, large, body, salt). The words overlap between both the sentence words S_I and S_JP.

Table 1: Sentence S_i, Extended-PMI implemented to teach min word in the sentence

Bodies	Tr = 0.39	Lay	Tr = 0.282	Boat	Tr = 0.421	Drown	Tr = 0.34	Sea	Tr = 0.34
Physical	0.39	Put	0.641	Small	0.455	Die	0.587	Division	0.607
Structure	0.451	Down	0.425	Vessel	0.488	Learned	0.331	Ocean	0.873
Animal	0.403	Coast	0.326	Travel	0.439	Inhalation	0.503	Large	0.738
Human	0.441	Area	0.336	Water	0.782	Water	0.634	Body	0.648
Related	0.436	Land	0.311	Tractor	0.362	Liquid	0.513	Salt	0.688
Adjectives	0.450	Sea	0.391	Reel	0.338	Submersion	0.000	House	0.173
Corporeal	0.407	Line	0.238	Skiff	0.361	Stopped	0.199	Upper	0.134
Identities	0.339	Veto	0.069	Cruiser	0.379	Pearsall	0.301	Direction	0.122
Weapons	0.344	Inside	0.137	Rowing	0.385	Kling	0.276	Central	0.052
Killed	0.342	Eyes	0.119	Paddle	0.316	Mull	0.140	Rain	0.167
Found	0.333	Suicide	0.119	Hopei	0.337	Spell	0.326	Hospital	0.118

Table 2: Sentence S_J, Extended-PMI implemented to teach min word in the sentence

Remains	Tr = 0.193	Floating	Tr = 0.148	Shore	Tr = 0.258	sinking	Tr = 0.287
Dead	0.444	buoyed	0.047	Fringe	0.425	Slow	0.639
Body	0.403	Water	0.387	Edge	0.264	Fall	0.667
Human	0.283	Liquid	0.567	Large	0.566	Decline	0.512
open	0.124	Shore	0.123	Body	0.352	Liquid	0.332
FOUND	0.109	Fringe	0.033	Water	0.370	plummeting	0.099
HUMAN	0.191	Edge	0.141	Island	0.110	Torpedoing	0.245
PLACE	0.120	Large	0.107	Ocean	0.232	Collapse	0.227
MYSTERY	0.143	Body	0.026	South	0.121	Plunged	0.177
ISSUE	0.169	Water	0.126	Chester	0.048	Slipping	0.258
Fallen	0.116	Plants	0.067	Harbor	0.273	Capsize	0.000
Tears	0.035	Voters	0.000	Hills	0.076	Falling	0.000

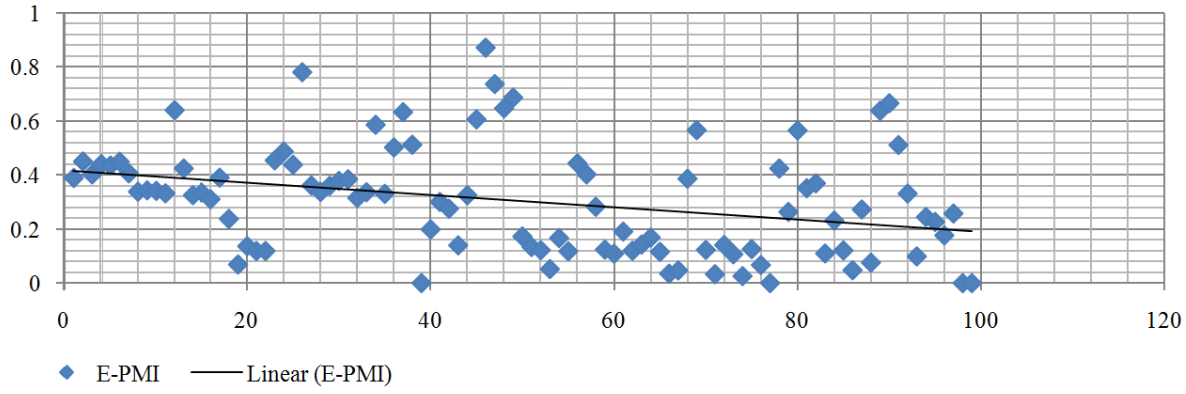


Fig. 2: Data distribution from the examples in Tables 1 and 2 and the relatedness of the cut-off value over all the E-PMIs

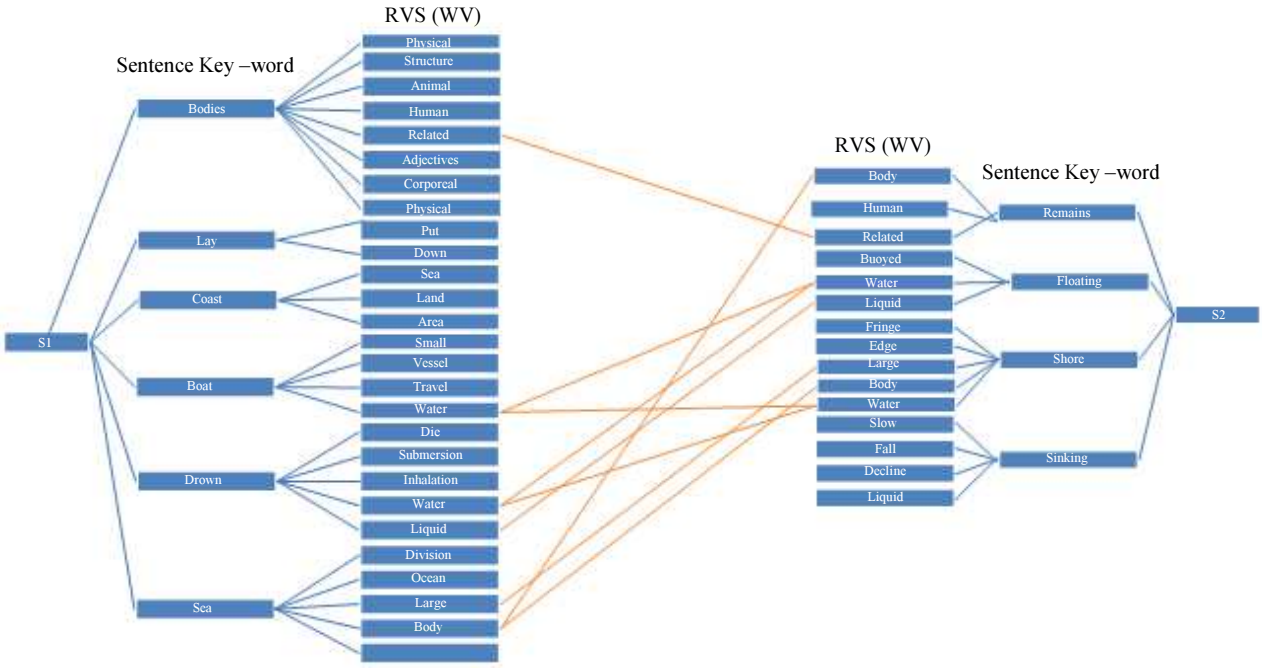


Fig. 3: Extracted word coverage overlap between short sentences, S_I and S_J

In Fig. 2, the semantic relatedness space appears in the whole example and the sparseness issue is also noted at the average of the semantic sparseness thresholds, where the linear Extended-PMI with its threshold is shown to have a sparseness that is lower than 0.355 and starts to be far from the semantic relatedness space of the RVS in both the sentences, S_i and S_j. Both S_i and S_J can be presented as in the following Fig. 3.

However, the similarity between $S_I$and $S_J$is very high, as can be seen in Fig. 3, by counting the number of shared RVS, where there are only 13 shared words providing a similarity of only 29% between the RVS of the words. The important words which present the meaning of the sentence share no similarity with the RVS. Here, the coverage limitation can be noticed by the lack of sufficient semantic relations between both words. In order to maintain this gap, the study was aimed at accomplishing two contributions: First, an increase in the RVS of the $S_I$and S_J' words through the ontology by determining the Most Related Gloss (MRG). A structural relationship between the RVS ofS_I and S_J is defined by measuring the nearest set of gloss word definitions.

Semantic Structural Information Content (SSI) Method based on Extracted Wiki-Gloss

The SSI method measures the semantic structural space based on the RVS. In this step, each RVS word is used as afulcrum through the gloss tree definitions

to determine the structural information in the WordNet. Each gloss includes a set of words, the ICGR, which defines the structural information of each gloss keyword (WG) regarding the RVS words for S_I and S_J to determine the Most Relatedness Gloss (MRG). Each RVS includes a set of gloss words with a full definition.

Each relatedness vector set includes an RVS set that presents the relatedness of Wiki to each word in $S_I$and S_J. Each RVS includes a set of gloss words with a full definition.For each WS word in the S_I, there is a set of words, WV, presented in the RVS which is considered as having the most relatedness from the previous step, $WS = WV_1, WV_2 ..... WV_N$.For each$WV$, G is extracted from WordNet as follows:

$$Gloss\, S_I\left(WV\right) = \forall \mid Gloss(G) \tag{7}$$

In this step, all the glosses of a particular RVS (WV) are defined and extracted. Each (WV) in the RVS has hundreds and thousands of glosses extracted from WordNet, but to reduce the gloss space, only the gloss that includes the main words in S_I and S_J are presented as the RVS-gloss as follows:

$$Gloss\left(RG\right) = \left(SW, \forall \mid Gloss(RVS\right) \tag{8}$$

The WS presents words that are used to search the gloss space, where only the vectors that contain the WS are extracted with Equation 8. Consider the words “Coast” and “Shore”, which share the same character in the sentence. The Extended-PMI is unable to introduce information that fits to fill the gap and define the similarity between the RVS. In the following section, a new Structural Semantic Information (SSI) method is introduced based on the use of a modified path with a local information distance measurement between the gloss keywords (GW) and W-wiki vectors for each gloss in the taxonomy. Then, the most related gloss to the WV is found. Next, the relationship between the words in the gloss and the W-wiki is defined based on the following definitions, which present the SSI as the core of the SSI.

Definition 1

The global dimension of the entire gloss taxonomy between the WG and WV is defined by the total information on path and depth and is normalized with the total gloss taxonomy tree. Similar methods use the depth of two concepts by finding the maximum depth, MaxDepth(C), among the two concepts, C1 and C2, which cannot define the relationship between the gloss taxonomy and the whole word gloss. In the case of this study, the gloss taxonomy tree is defined as in the following Fig. 4.

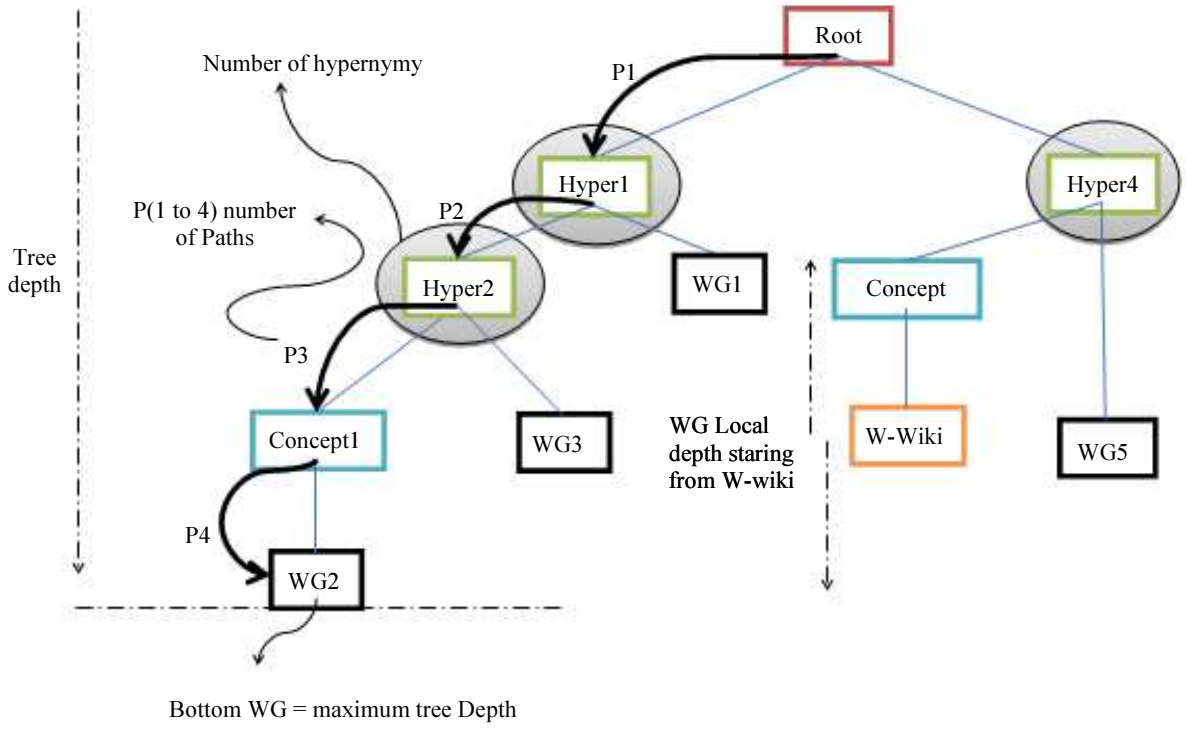


Fig. 4: Defining the most related gloss with the Wiki-word semantic by providing structural information from the WordNet taxonomy

The RVS (WV) and gloss words (WG) of the Global Semantic Structural Information (GSSI (WG_N)) is measured as in Equation 9:

$$GSSI\left(WG_N\right)=\frac{MaxMaxDepth\left(WG\right)+\left(\frac{1}{2*\left(MaxMaxPath-Min\ Pth\left(WG,WV\right)\right)}\right)}{Maximum\ G\ Depth} \quad (9)$$

where, (MaxDepth(WG)) represents the maximum depth and*WG*' bottomis WG2 in the gloss taxonomy tree in Fig. 4. The path is defined as twice the variance of the maximum and minimum paths (MaxPath-MinPath) of each *WV* and *WG* with the taxonomy root. The maximum gloss depth (Maximum G Depth) normalized all the arguments.

Definition 2

This is the local information needed to define the local depth of each *WG* to the *WS*, where the RVS (*WV*) is the settled point. The number of branches (hyper) required to move from the WV branch to the *WG* branch is defined as the number of hypernyms between both words. The Local Semantic Structural Information of gloss words LSSI(WG_N) is determined as in Equation 10:

$$\left(LSSI\left(WG_N\right)\right)=\frac{1}{Local\ depth\left(WG,WV\right)*Number\ of\ Hypernym\left(WG,WV\right)} \quad (10)$$

where, the local depth variation between the *VW* and the *WG* is multiplied with the number of sub-tree heads in the taxonomy (a hypernym is a generalisation of a synset). The local depth is 1 if *WV* and *WG* are at the same level in the taxonomy and it increases with each level between both the *WV* and *WG*. This is to ensure that both words in the same tree have an almost similar or higher value. In other words, if *WV* and *WG* are in different branches of a gloss tree taxonomy, then, a number of hypernyms must be used to bridge this gap, where, if *WG* and *WV* were in the same branch, the number of hypernyms would be (*WG*, *WV*) =1. Otherwise, the number of hypernyms would increase by 1 for each node that is counted. This is aimed at normalizing the *WG* and *WV* that have a similar depth but are in different trees in order to obtain a fair score.

Definition 3

This is the Global Semantic Structural Information (GSSI) that is normalized with the Local Semantic Structural Information (LSSI) between the WG and WV within the gloss taxonomy tree.

$$SSI\left(WG_N\right)=GSSI\left(WG_N\right)*LSSI\left(WG_N\right) \quad (11)$$

In this case, the *GSSI* is normalised with the*LSSI*, where if the *LSSI* value is equal to 1, it means the *WV* and *WG* are in the same tree and at the same local depth. The semantic relatedness was previously defined for each word in the gloss with the W-Wiki. As a final step, the *ICGR* is determined with the defined relatedness of all the glosses to the W-wiki as follows:

$$ICGR\left(G_I\right)=\frac{\sum_{\left(WG_n\right)\in G_I}SSI\left(WG_N\right)}{N} \quad (12)$$

where, WG_N represents the set of words in the gloss (G_I). The total SSI for the gloss (G_I) is divided by the number of words (N) in the gloss taxonomy tree. For each *RVS*, every gloss has a set, *SSI*(G_I), which shows the gloss density surrounding the *WV* in the taxonomy. The particular *WV* of the *RVS* includes a set of extracted MRG:

$$\left(WV\right)=\begin{pmatrix}ICGR\left(G1\right),ICGR\left(G2\right),\\ ICGR\left(G3\right),ICGR\left(G4\right)\ldots ICGR\left(G_I\right)\end{pmatrix}$$

where, each gloss has an ICGR value. The gloss is indexed in descending order and only the top five glosses are inducted into the relatedness matrix to expand the semantic space of the RVS by adding the words in these glosses to S_I and S_J.

By applying the SSI method, the number of glosses can be summed up as in Equation 12. The ICGR is presented as the normalized total number of words in the last row, which appears to find the most related set of *WG* to each *WV* (Water, sea). By measuring the ICGR to the vectors, in this example, there were two glosses for each vector from the two sentences to expand and wrap the gap between "Shore" and "coast". In this example, the *MRG* on the *WV* "Water" is the *SSI* (WG_1) that has been inserted into the *WS* sea-water in Fig. 3, where the word, "Water", is expanded to become (Unusual, destruct, rise, sea, shore, Storm, combination, wind, high and tide). Similarly, the *WV* "Sea" of the word Coast has two glosses, where $SSI(WG_2)_1$is the most related by expanding the related word of "Sea" with Coast, Sighted, Voyage, Flight and Water.

From Fig. 5, "Shore" and "Coast" are connected with two more words, which define the MRG to increase the semantic space similarity. The gap in Fig. 3 is maintained with two shared concepts between both the words "Shore" and "Coast". Only one gloss is used in our example and it still provides a good example of how the ICGR maintains the gap. The use of the top 5 MRG glosses to expand the semantic space of the Extended-PMI will give high relatedness points to both WV words.

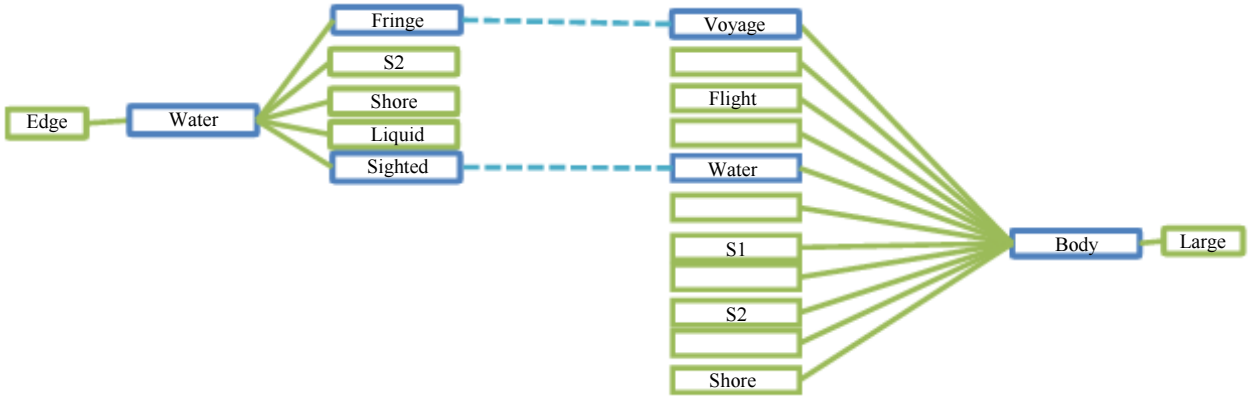


Fig. 5: Gap between Shore and Coast maintained with more than two GW-inducted Wiki spaces using just one gloss inducted from the ten glosses

Moreover, the MRG is not affected by the number of word variations in Fig. 5, where the regular rule is the more words, the higher the rank. The MRG is still able to detect the related gloss regardless of the number of words, as the best gloss (WG_2) has fewer words with 5 words and the other has double the number of words. These two advantages indicate the effectiveness of this method in finding the most related gloss that also contains two words that appeared in the Wiki word, "Water".

Finally, the words in the short text (SW) are enriched with regard to their WV extracted through the Extended-PMI using a Wikipedia-constructed corpus. Then, a high and well-constructed semantic space of related words is defined for each S_I and S_J by using the SSI to define the MRG between the G (WG) and RVS (WV). Finally, the proposed ICGR approach was proven to wrap the gap between the general concepts of *WV* and *WG* (Hypernym) that are near to the top of the hierarchy, appear to represent very broad distinctions and are presented as Nouns with concepts that are lower in the taxonomy representing a complex and very close similarity, even for concepts that are presented as Verbs.

ICGR Experimental Results

The ICGR approach was carried out in two steps. First, the Extended-PMI was used to extract the SVRs of each word in the short text. Second, the whole ICGR approach, which includes the Extended-PMI, was used with the SSI method to define the most relatedness gloss of each vector in the SVRs, where the semantic space of each word in the short text was expanded with words from the MRG. To define the relatedness similarity between pairs of words in the datasets in this evaluation, the word overlap strategy was adapted in this setting to define the similarity between the extended semantic spaces of the words in the datasets as the ICGR-overlap.

Evaluate Extended Pointwise Mutual Information (Extended-PMI) Method

The Extended-PMI approach is implemented to define the semantic vector relatedness through a search of pre-constructed corpora based on the Wikipedia dataset. The first step is to determine the training and testing of the cut-off threshold. Thus, a dataset with 178.6 MB of articles was selected from the Wikipedia dumps in November 20, 2017, where the bigram generated a high pair word space, which was used by the Extended-PMI as a search space to construct the highest Semantic Relatedness Vector (SRV) set.For example, a high bigram space will be ideal for a high SVR, which includes low density data. Thus, a cut-off threshold was tested to eliminate a low VRS frequency to W_i. As the Extended-PMI searches each word in the short text to define its SRV, this study considered a part of the WordSim-353 dataset as a training sample with 100 paired words.The size of the bigrams for each pair ranged from 65 to 300 words. However, only the top 100 bigrams appeared with each word that was considered. The following graph presents the size of the total dataset of bigrams, which is the percentage of shared words between the words of the total bigram size.

In this experiment, the Wikipedia articles were split into sentences. Then, bigrams were generated for each sentence, where the total number of bigrams generated was 1610000 pairs. Further, part of the WordSim-353 dataset of 100 pairs was used for the cut-off threshold training, where 70600 bigrams occurred with all the words in the dataset. The defining of the global threshold is an inaccurate strategy for determining the similarity due to the complexity of the natural language and unexpected changes, including the usage and reusability of a word in different contexts (MacCartney, 2009). In this experiment, only the top 100 words were tested further to find the best cut-off threshold, where the

Extended-PMI was implemented to 100 pairs to define the data sparseness point of the top 100 SRV and the average cut-off point. Experiments were carried out over 100 pairs (words from WordSim-353 and the words of the bigram that they appeared with). For example, a pair of words, "Drug" and "Abuse", was implemented to define their SRVs using the Extended-PMI to the top 100 bigrams. Figure 7 shows the data density for each vector and the leaner cut-off threshold based on their results, as in appendix E.

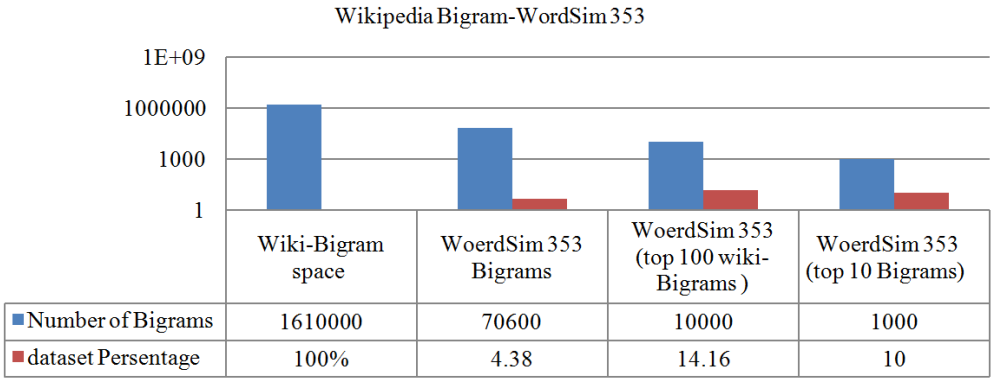


Fig. 6: Dataset selection sizes of Wikipedia bigrams

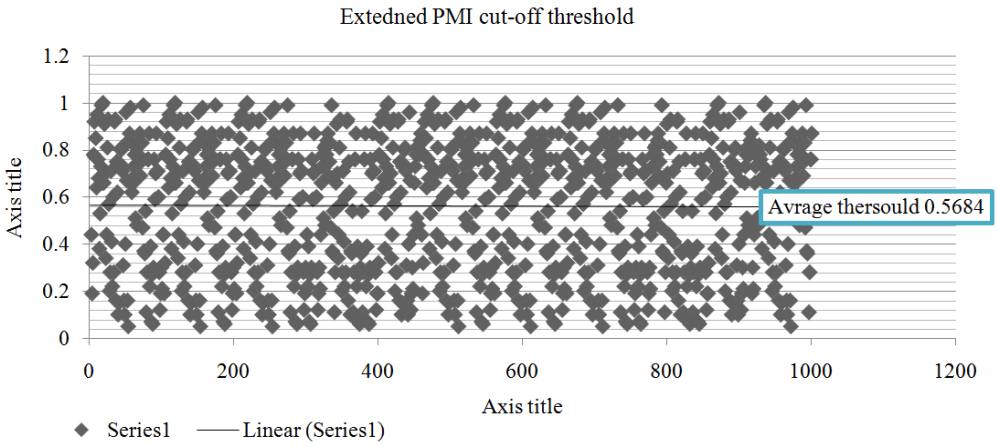


Fig. 7: Ten SRV densities for both the words "Drug" and "Abuse" in the data density threshold

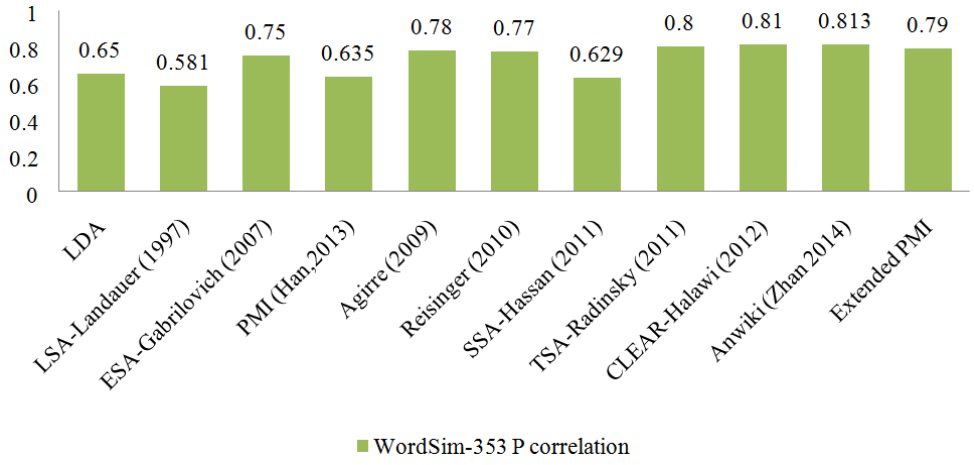


Fig. 8: Extended-PMI correlation compared to state-of-the-art methods

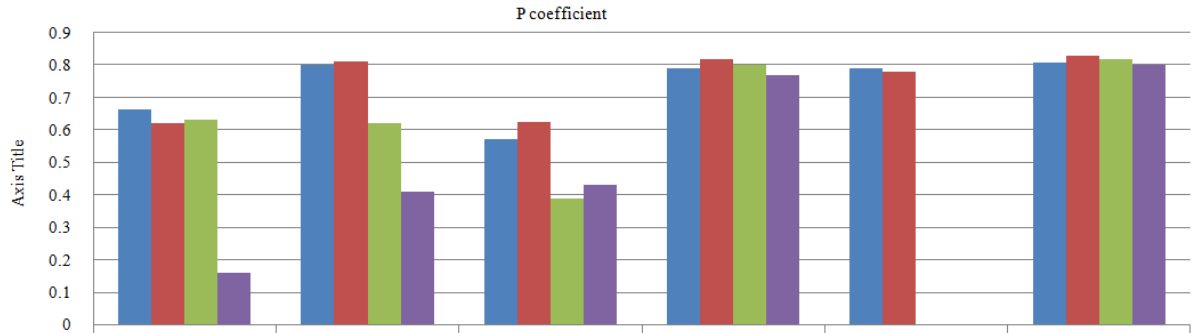
As the space of the semantic relatedness was high using the Extended-PMI, only the high-density vectors were considered as the most related vectors to each word in the dataset. Regarding the experimental results from the use of the WordSim-353 dataset in Appendix (E), the Pearson's coefficient was used to determine whether the proposed strategy could improve the word similarity relatedness through a high Wikipedia space by comparing it to a corpus set based on state-of-the-art methods, namely, ESA, LSA, PMI-WIK (Strube and Ponzetto, 2006), LDA, (Agirre *et al*., 2009; Gabrilovich and Markovitch, 2007; Luong *et al*., 2014; Radinsky *et al*., 2011) and the proposed Extended-PMI of this study. Figure 8 presents the correlation results as follows:

Figure 8, the proposed method showed a very good performance in terms of a Spearman's correlation (P) of 0.79 and it slightly outperformed two of the state-of-the-art methods, namely, CLEAR (Halawi *et al*., 2012) and TSA (Radinsky *et al*., 2011). The CLEAR method, which is based on a generalized vector space mode with WordNet (Halawi *et al*., 2012) had a correlation of 0.8. However, CLEAR with P=0.81 correlated with the human-based WOrdSim353 dataset, the method with the strictest and most limited semantic dimension space ranging from 100-150, which led to the elimination of the important concepts of the semantic space. CLEAR outperformed the original ESA method (Gabrilovich and Markovitch, 2007). The TSA method (Radinsky *et al*., 2011) with a correlation of 0.8, suffered from sparseness in certain topics, mostly specific topics such as technology and science, as presented in the experiments. The sparse issue, in addition to the strict SVD space, affected the dimensional space of the LSA (Landauer *et al*., 1997) and explained the low correlation to the human-based vector space method (Agirre *et al*., 2009; Reisinger and Mooney, 2010) to present the dimensional space of Wikipedia with correlations of 0.87 and 0.77, respectively. This was due to the data sparseness problem and even when using the concept of the window strategy by (Agirre *et al*., 2009) the data was still not large enough. The data sparseness issue continued to affect the results of the original PMI (Han *et al*., 2013) SSA (Hassan and Mihalcea, 2011) and LDA, which had correlations of 0.62 and 0.65, respectively. Zhang and Zhu (2015) presented the Wikipedia sentence level co-occurrences (SLC) (ANwiki (slc)), which shows limited performance by reverse-engineering the association strength into an aggregation function of a vector of structured co-occurrences. Also, Zhang relied on (Wettler and Rapp, 1993) to set a constant normalization vector for estimating a word association that only fitted the particular dataset used in their experiments. In the current dataset, this issue was slightly improved with good results by providing a trained cut-off threshold. However, the proposed method presented a very close result to that obtained by Zhang due to the parameter that was used to substantiate the general similarity result of 0.66 by Zhang, which met the determined threshold of 0.65 of this study, as presented in Fig. 9. Compared to the original PMI, the Extended-PMI reduced the sparseness issue by eliminating the low-density vectors in the SRV.

Evaluation of Information Content Glossary Relatedness (ICGR) Approach

The validity of the Extended-PMI was leveraged in the previous section. As the Semantic Structural Information (SSI) method was implemented for each RVS, the gloss relatedness was implemented as the ICGR-overlap between two words in the dataset to measure similar words in the semantic relatedness space based on Equation 8. In this section, the results of the experiments that were implemented using four datasets, as provided in appendices D, F, G and E, namely, RG-65, MC-30, MG-30 and WordSim-353, are presented. The Spearman's correlation coefficient (p) was used to compare the results of the MRG method with five gloss-based methods, namely LESK (Lesk, 1986), Gloss Vector (Patwardhan *et al*., 2003), EGO (Patwardhan and Pedersen, 2006), G2WS (Aouicha and Taieb, 2015) and G3W (Aouicha *et al*., 2016). Figure 9 shows the compressed results.

The ICGR method presented very good results for all the datasets. As the SSI was restricted to the shared words between the SRV and the main words, this led to a very good correlation with the human scores in all the four datasets. In the RG-65 and MC30 datasets, the ICGR slightly outperformed the reported results of the G2WS and the G3W by 2%. However, the GV had a very good result with 0.8 and 0.81% and this result was relayed only to the nouns detected in the dataset and ignored other word categories, which decreased in cases that included other gloss POSs. The gloss-overlap approaches, namely, LESK, GV and EGO, had an average correlation and significant drop in performance when just the glosses of the concept alone were use. In the GM30, the ICGR had the highest correlation to the human scores, outperforming all the state-of-the-art methods. In the WordSim-353 dataset, the performance of the LESK, GV and G2WS methods significantly decreased with correlations of 0.51, 0.21 and 0.3, respectively. The very low occurrence of overlapping shared words between the glosses failed to draw distinctions among the different glosses, whereas high-frequency words occurred in many glosses and again did not provide useful information to distinguish the relatedness among the glosses. Meanwhile, the ICGR was presented as having a stable performance among all the datasets, with an average difference of 2%.



	Lesk M (1986)	Gloss Vector (Patwardhan and Pedersen (2003))	Patwardhan and Pedersen (2006) EGO	G2WS (Aouicha, Mohamed Ben(2015)	G3W (Aouicha, Mohamed Ben (2016)	ICGR
RG65	0.662	0.8	0.571	0.79	0.79	0.81
MC30	0.62	0.813	0.623	0.82	0.78	0.83
GM 30	0.63	0.62	0.39	0.8	0	0.82
WordSim-353	0.158	0.41	0.43	0.77	0	0.8

Fig. 9: Spearman correlation of ICGR compared with gloss based approaches

Table 3: Results of PMI cut-off density threshold used in ICGR approach over four datasets

Dataset	ICGR cut off density threshold
RG65	0.43
MC30	0.23
GM 30	0.25
WordSim-353	0.56

The effective connection between the Wikipedia space and WordNet by presenting the SVR to define the MRG significantly increased the correlation of the proposed approach in all the semantic relatedness datasets. The use of SVRs to present the similarity by striking the GLOSS definition to the main word increased the correlation by up to 14-23%. This experiment used the same sized Wikipedia dataset with 178.6 MB from the 2017 Wikipedia articles, where Table 3 shows the cut-off threshold values.

Each dataset had its own threshold based on the part of the dataset that was used with the Extended-PMI. Three threshold values were released, where the large dataset had a high cut-off density threshold as in WordSim-353, where T= 0.56, while in the small datasets such as MC and GM30, the cut-off density threshold was 0.29 and 0.25, respectively. In each dataset, the best setting for the frequency cut-offs removed both the low-density vector words, thereby eliminating the source of noise in the used Wikipedia corpus. The very low frequency words did not occur enough to draw similarity relatedness between the words across Wikipedia and the WordNet space.

Evaluate ICGR Approach Short Text Similarity (STS) Task

In this section, the overall ICGR approach is evaluated to validate its capability to measure the short text similarity based on the semantic relatedness. The final step was the ICGR-similarity overlap as a measure of the semantic similarities between short texts. The dataset by Li *et al*. (2006) was used in these experiments. A set of challenges were posed by this dataset due to the high number of differences in the human evaluations presented in different works (a new word sense similarity measure in WordNet, a semantic text similarity using a corpus-based word similarity and string similarity and an analysis of word co-occurrences in human literature for supporting semantic correspondence discovery). Thus, a score of GT 65 was retrieved from the original dataset, which contained grades between 0-4. The scores were normalized by four to be used in the experiments. The dataset and the experiments are presented in Appendix (H). The R correlation coefficient results for the ICGR approach were compared to a set of state-of-the-art semantic relatedness methods, namely, LG (Lee *et al*., 2014)), SyMSS (Oliva *et al*., 2011), Li McLean (Li *et al*., 2006), STS-Islam (Islam and Inkpen, 2008), Omiotis (Tsatsaronis *et al*., 2010) (2010), LSA (O'Shea *et al*., 2008), LIU (Liu *et al*., 2007), LDA (Guo and Diab, 2012)and ANwiki(Zhang and Zhu, 2015).

In Fig. 10, the results were divided into four parts, where ANwiki and ICGR presented the highest correlation with 0.94 and 0.89, respectively. The ANwiki approach took advantage of the existing high number of features through the Wikipedia space. Omiotis and SOC-PMI achieved a similar correlation of 0.85 to the human scores, where the correlation decreased between 5-9% to the first two approaches. LDA, LIU, LSA, LI McLean and SyMSS achieved a correlation of 0.84-0.8 to the human scores, where the correlation decreased to 10% lower than the top approaches. The LG approach had the lowest correlation to the human scores on this dataset, where the correlation decreased to 30%.

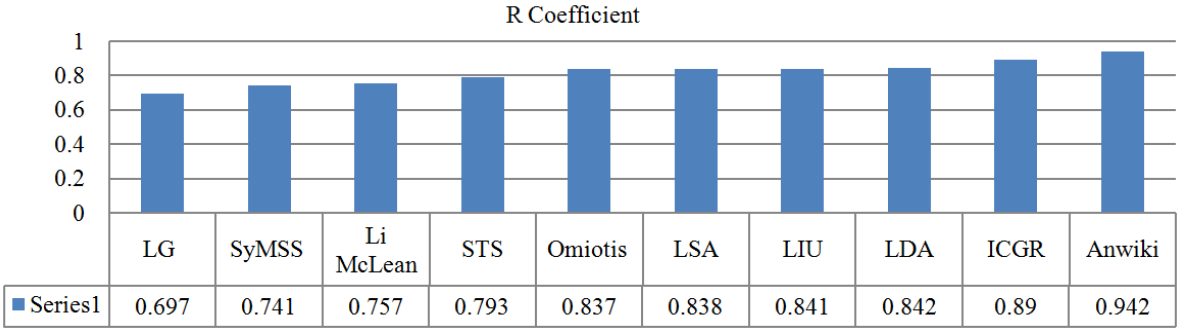


Fig. 10: Comparison of ICGR approach to a set of state-of-the-art methods

The high semantic space constructed by ANwiki produced high semantic features, which carried extensive information. However, the approach was restricted to the training of free association norms, which considered the human notation and evaluation of a particular dataset. This fact affected the performance of the approach since no results were associated to the dataset. However, the correlation of the SOC-PMI approach decreased according to the dataset size, where the limited size of the dataset with same meanings was expressed by different syntactic orders, which were not observed in the short texts. LIU and Omiotis, followed the same strategy of defining the shortest path across all the paths between two words besides the limited depth measurement based on the standard subsume, which only considered the shared subsume depth and did not define the whole tree depth. However, for both the LDA and LSA, the sizes of the words by the context matrix were limited and fixed to several hundred. As a result, the vector was fixed and thus, was likely to be a very sparse representation of a short text such as a sentence. The word-by-word matrix did not capture the meaning of the sentence well and the sentence vector became diluted as a large number of words were added to it (Li *et al*. 2006). The fourth set of methods were the LG, SyMSS and Li McLean, which shared a similar methodology of combining syntactic and semantic features to measure the short text similarity. These methods shared the same strategy of combining semantic and syntactic features to measure the short text semantic similarity. These methods achieved a low correlation due to the use of standard semantic similarity measurements, such as WUP, along with syntactic features.

The proposed ICGR approach outperformed at this point by contributing very well at each validation stage. The ICGR defined the whole gloss depth along with the number of transactions between both the Wikipedia vectors and the MRGs of the words. For example, the LSA and SOC-PMI presented a very low correlation of 0.45 and 0.6 for “Shore” and “Coast”, respectively to the human notation of 0.92. The MRG was restricted to the main words of the short texts to eliminate the sparseness issue, thereby decreasing the correlations of the LSA, DLA and SOC-PMI. A set of gloss keywords was defined for each main word in the short text, where a total of 346 words were associated with “Shore” and a total of 430 words were associated with “Coast” out of a total of 38 glosses, where the ICGR overlap achieved a score of 0.78, which was a good result compared to the human score of 0.9. However, the high semantic space generated by the ICGR approach was not free of abuse. Words with similar meaning but not related to the context affected some of the correlation results, thereby working to decrease the correlation. The following section presents the integrated weighted grammatical relations extracted by the DSW along with the ICGR approach.

Conclusion

In this study, the ICGR increased the semantic space with robust semantic words. The Extended-PMI, based on a cut-off density threshold, was aimed at decreasing the unrelated search space in the generated corpus by defining a reliable threshold of the Extended-PMI for each word in the short text. Only the top ten RVS were extracted. Then, the SSI method was used to define the gloss of each vector defined to each Wikipedia vector by restricting the search to the main words of the short text. In other words, only the glosses that were related to the Wiki vectors and that included the main words of the short text were considered, which led to a decrease in the sparseness issue in most of the information content methods. The SSI method defined the gloss relatedness over the whole definition instead of the regular word overlap using the RVS. The experimental results were implemented to the Extended-PMI using WordSim353, which achieved a very good correlation to the human scores. Meanwhile, the ICGR approach achieved very good results compared to the human scores for the benchmark datasets.

Besides the semantic features, the grammatical relations also contributed to the understanding of natural language sentences. Thus, in future studies, a grammatical relation extraction is proposed to extract and enrich the

grammatical relations of limited contexts and to provide an integrative strategy to measure short text similarities based on grammatical and semantic features.

Author's Contribution

Ali Muftah BenOmran: Participated in all experiments such as data preparation, selection and testing coordinated the data-analysis and contributed to the writing of the manuscript.

Mohd Juzaiddin Ab Aziz: Designed the research plan, organized the study, contribute in revision of manuscript, participated in all experiments, coordinated the data-analysis and contributed to the review of the manuscript.

Ethic

No ethical issues would arise after the publication of this manuscript.

References

Al-Baltah, I.A., A.A.A. Ghani, W.N.W. Wan Nurhayati and R. Atan, 2014. Semantic conflicts detection of heterogeneous messages of web services: Challenges and solution. J. Comput. Sci., 10: 1428-1439. DOI: 10.3844/jcssp.2014.1428.1439

Alzahrani, S., 2016. Cross-language semantic similarity of Arabic-English short phrases and sentences. J. Comput. Sci., 12: 1-18. DOI: 10.3844/jcssp.2016.1.18

Aouicha, M.B., M.A. Hadj Taieb and A.B. Hamadou, 2016. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Applied Intell., 45: 475-511. DOI: 10.1007/s10489-015-0755-x

Aouicha, M.B. and M.A.H. Taieb, 2015. G2WS: Gloss-based WordNet and Wiktionary semantic Similarity measure.

Banerjee, S. and T. Pedersen, 2003. The design, implementation and use of the Ngram statistics package. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing, Feb. 16-22, Springer, Mexico City, Mexico, pp: 370-381. DOI: 10.1007/3-540-36456-0_38

BenOmran, A. and M.J.A. Aziz, 2014. Syntactically enhanced LSA methods in automatic essay grading systems for short answers. Proceedings of the 3rd International Conference on Computer Engineering and Mathematical Sciences (EMS' 14), At Langkawi, Malaysia, pp: 412-417.

Bollegala, D., Y. Yoshida and K. Kawarabayashi, 2017. Using k-way Co-occurrences for Learning Word Embeddings. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, Feb. 02-07, New Orleans, Louisiana, USA, pp: 1-17.

Chen, P., W. Ding, C. Bowes and D. Brown, 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. Computational Linguistics. DOI: 10.3115/1620754.1620759

De Luca, E.W. and A. Nürnberger, 2006. Using clustering methods to improve ontology-based query term disambiguation. Int. J. Intell. Syst., 21: 693-709. DOI: 10.1002/int.20155

Firth, J.R., 1957. Vector-space semantics (distributional similarities) Distributional similarities Distributional similarities (Firth).

Gabrilovich, E. and S. Markovitch, 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07), pp: 1606-1611. DOI: 10.1145/2063576.2063865

Guo, W. and M. Diab, 2012. Modeling sentences in the latent space. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Jul. 08-14, Association for Computational Linguistics, Jeju Island, Korea, pp: 864-872.

Halawi, G., G. Dror, E. Gabrilovich and Y. Koren, 2012. Large-Scale Learning of Word Relatedness with Constraints, pp: 1406-1414.

Han, L., T. Finin, P. Mcnamee, A. Joshi and Y. Yesha *et al*., 2013. Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy. EEE Trans. Knowl. Data Eng., 25: 1307-1322.

Hassan, S. and R. Mihalcea, 2011. Semantic relatedness using salient semantic analysis. Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011), pp: 884-889.

Hirst, G., 2006. Distributional measures of semantic distance: A survey.

Islam, A. and D. Inkpen, 2008. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discovery Data, 2: 1-25. DOI: 10.1145/1376815.1376819

Jabeen, S., 2014. Exploiting wikipedia semantics for computing word associations. Victoria University of Wellington.

Jaccard, P., 1982. Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientales. Bull. Murithienne, 37: 81-92.

Jager, B., M.J. Green and A.A. Cleland, 2016. Polysemy in the mental lexicon: Relatedness and frequency affect representational overlap. Language Cognition Neurosci., 31: 425-429. DOI: 10.1080/23273798.2015.1105986

Javanmardi, S. and C. Lopes, 2010. Statistical measure of quality in Wikipedia. Proceedings of the 1st Workshop on Social Media Analytics,Jul. 25-28, ACM, Washington D.C., District of Columbia, pp: 132-138. DOI: 10.1145/1964858.1964876

Jimeno-Yepes, A. and A.R. Aronson, 2012. Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, Jan. 28-30, ACM, Miami, Florida, USA, pp: 733-736. DOI: 10.1145/2110363.2110449

Kilgarriff, A. and J. Rosenzweig, 2000. Framework and results for English SENSEVAL. Comput. Humanities, 34: 15-48. DOI: 10.1023/A:1002693207386

Landauer, T.K., D. Laham, B. Rehder and M.E. Schreiner, 1997. How well can passage meaning be derived without using word order? A Comparison of Latent Semantic Analysis and Humans.

Lee, M.C., J.W. Chang and T.C. Hsieh, 2014. A grammar-based semantic similarity algorithm for natural language sentences. Scientific World J., 2014: 437162-437162. DOI: 10.1155/2014/437162

Lesk, M., 1986. Automatic sense disambiguation using machine readable dictionaries. Proceedings of the 5th Annual International Conference on Systems Documentation, (CSD' 86), Toronto, Ontario, Canada,ACM, pp: 24-26. DOI: 10.1145/318723.318728

Li, Y., D. McLean, Z.A. Bandar, J.D. O'Shea and K. Crockett, 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng., 18: 1138-1150. DOI: 10.1109/TKDE.2006.130

Liu, X., Y. Zhou and R. Zheng, 2007. Sentence similarity based on dynamic time warping. Proceedings of the International Conference on Semantic Computing, Sept.17-19, IEEE Xplore Press, Irvine, CA, USA, pp: 250-256. DOI: 10.1109/ICSC.2007.48

Lucie, I., 2017. Leveraging lexical-semantic knowledge for text classification tasks. Technische Universität.

Luong, N.Q., L. Besacier and B. Lecouteux, 2014. Some Propositions to Improve the Prediction Capability of Word Confidence Estimation for Machine Translation. J. Comput. Sci. Communi. Eng., 30: 36-49.

MacCartney, B., 2009. Natural language inference. Stanford University.

O'Shea, J., Z. Bandar, K. Crockett and D. McLean, 2008. Pilot short text semantic similarity benchmark data set: Full listing and description. This Technical Report has been Submitted to Turnitin.com.

Oliva, J., J.I. Serrano, M.D. del Castillo and Á. Iglesias, 2011. SyMSS: A syntax-based measure for short-text semantic similarity. Data Knowl. Eng., 70: 390-405. DOI: 10.1016/j.datak.2011.01.002

Patwardhan, S., S. Banerjee and T. Pedersen, 2003. Using measures of semantic relatedness for word sense disambiguation. Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, (TPCL'03), pp: 241-257. DOI: 10.1007/3-540-36456-0_24

Patwardhan, S. and T. Pedersen, 2006. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, (ACL'06) pp: 1–8.

Pesaranghader, A., S. Muthaiyah and A. Pesaranghader, 2015. Improving gloss vector semantic relatedness measure by integrating point wise mutual information. Proceedings of the International Conference on Informatics and Creative Multimedia, Sept.4-6, IEEE Xplore Press, Kuala Lumpur, Malaysia, pp: 196-201. DOI: 10.1109/ICICM.2013.41

Petrakis, E.G.M. and G. Varelas, 2006. X-Similarity: Computing semantic similarity between concepts from different ontologies.

Radinsky, K., E. Agichtein, E. Gabrilovich and S. Markovitch, 2011. A Word at a Time□: Computing Word Relatedness using Temporal Semantic Analysis. In Proceedings of the 20th International Conference on World Wide Web (WWW' 11), pp: 337-346.

Rodríguez, M.A. and M.J. Egenhofer, 2003. Determining semantic similarity among entity classes from different ontologies. IEEE Trans. Knowl. Data Eng., 15: 442-456. DOI: 10.1109/TKDE.2003.1185844

Reisinger, J. and R.J. Mooney, 2010. Multi-Prototype Vector-Space Models of Word Meaning, pp: 109-117.

Strube, M. and S.P. Ponzetto, 2006. WikiRelate! Computing semantic relatedness using Wikipedia. Proceedings National Conference Artificial Intelligence, 21: 1419. DOI: 10.1.1.231.9545

Sweta, D. and M. Singh, 2017. Improving question answering system based on a hybrid technique. J. Comput. Sci. DOI: 10.3844/ofsp.11390

Tsatsaronis, G., I. Varlamis and M. Vazirgiannis, 2010. Text relatedness based on a word thesaurus. J. Artificial Intell. Res., 37: 1-39. DOI: 10.1613/jair.2880

Tversky, A., 1977. Features of similarity. Psychol. Rev., 84: 327-352. DOI: 10.1037/0033-295X.84.4.327

Wettler, M. and R. Rapp, 1993. Computation of Word associations based on the co-occurrences of words in large Corpora. Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives, (AIP' 93), pp: 84-93.

Zesch, T., I. Gurevych and T. Zesch, 2016. Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words.Nat. Lang. Eng., 16: 25-59. DOI: 10.1017/S1351324909990167

Zhang, K. and K.Q. Zhu, 2015. An association network for computing semantic relatedness.Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, (CAI' 15), Association for the Advancement of Artificial, pp: 593-599.