

Article

A Method for Perception and Assessment of Semantic Textual Similarities in English

Omar Zatarain ^{1,*}, Jesse Yoe Rumbo-Morales ¹, Silvia Ramos-Cabral ¹, Gerardo Ortiz-Torres ¹, Felipe d. J. Sorcia-Vázquez ¹, Iván Guillén-Escamilla ² and Juan Carlos Mixteco-Sánchez ²

¹ Department of Computer Science and Engineering, CUValles, University of Guadalajara, Guadalajara 46600, Jalisco, Mexico; felipe.sorcia@academicos.udg.mx (F.d.J.S.-V.)

² Department of Natural and Exact Sciences, CUValles, University of Guadalajara, Carr. Guadalajara-Ameca Km. 45.5 Ameca, Guadalajara 46600, Jalisco, Mexico

* Correspondence: omar.zatarain@academicos.udg.mx

Abstract: This research proposes a method for the detection of semantic similarities in text snippets; the method achieves an unsupervised extraction and comparison of semantic information by mimicking skills for the identification of clauses and possible verb conjugations, the selection of the most accurate organization of the parts of speech, and similarity analysis by a direct comparison on the parts of speech from a pair of text snippets. The method for the extraction of the parts of speech in each text exploits a knowledge base structured as a dictionary and a thesaurus to identify the possible labels of each word and its synonyms. The method consists of the processes of perception, debiasing, reasoning and assessment. The perception module decomposes the text into blocks of information focused on the elicitation of the parts of speech. The debiasing module reorganizes the blocks of information due to the biases that may be produced in the previous perception. The reasoning module finds the similarities between blocks from two texts through analyses of similarities on synonymy, morphological properties, and the relative position of similar concepts within the texts. The assessment generates a judgement on the output produced by the reasoning as the averaged similarity assessment obtained from the parts of speech similarities of blocks. The proposed method is implemented on an English language version to exploit a knowledge base in English for the extraction of the similarities and differences of texts. The system implements a set of syntactic and logical rules that enable the autonomous reasoning that uses a knowledge base regardless of the concepts and knowledge domains of the latter. A system developed with the proposed method is tested on the “test” dataset used on the SemEval 2017 competition on seven knowledge bases compiled from six dictionaries and two thesauruses. The results indicate that the performance of the method increases as the degree of completeness of concepts and their relations increase, and the Pearson correlation for the most accurate knowledge base is 77%.

Keywords: knowledge bases; unsupervised reasoning; concept elicitation; computational epistemic skills; semantic text similarity

MSC: 68W32



Citation: Zatarain, O.; Rumbo-Morales, J.Y.; Ramos-Cabral, S.; Ortiz-Torres, G.; Sorcia-Vázquez, F.d.J.; Guillén-Escamilla, I.; Mixteco-Sánchez, J.C. A Method for Perception and Assessment of Semantic Textual Similarities in English. *Mathematics* **2023**, *11*, 2700. <https://doi.org/10.3390/math11122700>

Academic Editors: Laura Cornejo-Bueno and Jorge Pérez-Aracil

Received: 9 March 2023

Revised: 11 April 2023

Accepted: 18 April 2023

Published: 14 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Semantic text similarity is a challenging computational task due to linguistic issues of each natural language such as the use of synonyms, polysemous words and phrases, named entity recognition, multiple syntax rules that can be used to produce a single semantics, unbounded expression styles of writers, obscure semantics, the presence of beliefs, aphorisms, and subtle ways to measure the content within the texts. These issues may be addressed with complex algorithms, or they may be solved using deep learning, machine learning artificial intelligence, statistics, probabilities, and belief networks with good-to-excellent

success on the memorization of reasoning on past knowledge. However, empirical evidence and psychological studies on human learning suggest that humans develop behaviors focused on the perception of language and as a product of language acquisition, for example, a person produces judgments on the perceived information using their own knowledge and/or by using several sources to contrast the evidence. A natural language as a vehicle to transfer information has a grammar and lexicon that are used in a context-free structure. The semantics of a text is defined freely according the autonomous will of the writer. Our proposed work focuses on the exploitation of a knowledge base regardless of its content and makes no assumptions on the latter. The content of a pair of text snippets must be processed by the exploitation of the knowledge base to detect the similarities. Therefore, it is desirable that the machines exhibit intelligent behaviors by mimicking basic skills of perception, reasoning and judgement of semantic text similarities. Keeping in mind that many information sources have unstructured data and most unsupervised methods require big amounts of labeled data, our research is motivated on how machines can produce their own learning given a text and one or several knowledge bases, as well as how a pair of text snippets may be converted into related/unrelated concepts by mimicking a few behaviors of analysis and reasoning.

The first question requires skills for the extraction of information and disambiguation processes that identify the parts of speech in a source of information without labels assigned in advance; thus, it is necessary to implement a disambiguation process that identifies the role of a word by the relative position within the text. The second question requires strategies that emulate the detection of similar semantics related first to a pair of words and, ultimately, the semantics of a pair of sentences regarding their parts of speech.

This research develops a method that emulates the perception of the parts of speech, detects the biases produced in the early perception to restructure the main parts of a sentence, applies a reasoning process consisting of the detection of similarities due to several rules that describe the possible linguistic scenarios and finally obtains an assessment of similarity on the findings. The method is proposed for: (1) perception from scratch of the parts of speech of texts through exploitation of knowledge bases, (2) mimicking a few basic skills for detection of biases on the perceived parts of speech, (3) reasoning about the types of similarities and producing assessment judgements on semantic similarities.

The rest of this work is structured as follows: Section 2 mentions the state of the art regarding semantic text similarity, Section 3 describes the proposed method to autonomously perceive the parts of speech, the correction on the perception or debiasing, the process of reasoning about the combinations of the text pairs and the criteria for assessing a numeric similarity judgement, Section 4 defines several experiments on semantic textual similarity to observe the performance of the algorithms produced using the proposed architecture, Section 5 describes the outcomes of the performed experiments, Section analyses the results of the experiments and the differences between the proposed solution and the state of the art, and finally, Section 6 provides the conclusions and future work.

2. State of the Art

A learning theory based on the proximal development zone of Vigotsky [1] suggests that human beings produce their own individual knowledge and skills by the exposure to information and stimuli close to the individual, and the learning is produced by the interactions of the learners with the environment. On the other hand, Skinner [2] proposes behaviorism where learning is based on the reinforcement and repetition of behaviors. The two approaches have raised the interest of exploring new ways to enable machines which learn about knowledge, methods that focus on interaction with the environment and reinforced learning, and classical artificial intelligence such as rule-based reasoning. An example of a semi-supervised method that elicits concepts [3] by the extraction of common knowledge from a set of related definitions to produce the social knowledge of such a concept can be found in [4]. Knowledge bases [5,6] are sources of information of structured (labeled) or unstructured (unlabeled) data that have been used to extract

concepts and find the common information shared by definitions related to a single concept autonomously with the aid of basic similarity functions such as concept elicitation. A psycho linguistic contribution is the Wordnet for knowledge representation [7], enabling a few unsupervised techniques for knowledge acquisition [5]. Methods that focus on behaviors are supervised methods where a system is trained to remember facts using a predefined set of classes. One supervised strategy used on semantic text similarity is recurrent neural networks [8]; a few examples that use deep neural methods for semantic similarity are the Gaussian mixture model [9], the universal model for multilingual and cross-lingual semantic text similarity [10] and the semantic information space [11]. The Gaussian mixture model (GMM) [9] combines alignments and sentence-level embeddings, using the Hungarian algorithm [12] for word alignment, word-to-word similarity [13] and Wordnet [7] for assessment of word similarity. This method is inspired by a previous GMM used for open-ended questions [14] to represent feature vectors and compute the membership weights of semantic levels. The universal model for multilingual and cross-lingual semantic text similarity [10] combines word embedding [13] with N-gram overlaps with support vector regression [15], longest common prefix, use of kernel spaces for common structures [16], alignment features [17], a sequential process of deep averaging networks [18] and long short-term memory neural networks [8]. The labeling and part of speech is aided by the Stanford CoreNLP [19], which is also a supervised strategy. Semantic information space [11] uses taxonomies and the Jaccard similarity with word alignments [17] to compute information content [20] with support vector machines [21]. Transformers [22] are a type of neural networks that enable the pre-training and fine tuning of a family of BERT methods that use language-masked models to predict the context by masking tokens. The first method known as BERT [23] is a type of bidirectional network that uses encoder representation; the network architecture used in the method is applied to both the pre-training and the fine tuning and one advantage of this is that it can perform diverse nlp tasks. This method requires 110M and 340M parameters in the base and large versions, respectively. The pre-training data were 800M words and 2500M words to apply a prediction of masked tokens and a next-sentence prediction. An interesting approach developed by this method was the self-attention thanks to the bidirectional training. Due to the number of parameters in BERT being big, a number of strategies focus on reducing the size of the model. One of them is ALBERT [24]. This method reduces its sizes from 235M to 12M parameters and improves the results on different datasets when the parameters are shared across the layers. The results improve by 93% when trained with 1.5M of steps, and the training time is reduced to 32 out of 34 h compared to BERT. Sentence BERT [25] is a strategy that uses Siamese and triplet BERT networks, cosine similarity and is trained using a classification objective function, regression objective function and a triplet objective function. The latest state of the art on semantic textual similarity has focused on the development of BERT models by improving the training of these models; their performances vary from 71% to 93% on the Pearson correlation metric [26–30].

Despite the success of deep learning strategies mentioned above, the constraints of the need of training on big amounts of data and the assumption of a set of classes on the data prevent learning when new samples are out of training data.

Unsupervised methods on semantic text similarity comprehend the use of synonym sets (synsets) [31] by exploiting Babelnet [5] and various types of alignments on strings using kernel functions and the Stanford CoreNLP toolkit for pre-processing [19]. This method uses word embedding to compute the similarity of two sentences with vectors of 400 dimensions, a soft cardinality measurement of non-identical elements within a set, a weighted aligner and a dissimilarity computation known as edit distance based on the Levenshtein distance of words. Another unsupervised method applied on multilingual semantic text similarity [32] defines paragraph vectors [13] that are trained with independence of the knowledge content and uses three similarity metrics: cosine, Bray–Curtis and correlation.

3. Method for Semantic Textual Analysis

3.1. Notation Used in This Work

The following notation is used in the proposed method and architecture:

- T: a text snippet that may contain either a non phrase, sentence or question.
- NP: a noun phrase consisting either by a pronoun, (adjective)* noun, or entity.
- Q: a question consisting either by a structure denoting a question ending with a question mark (?), or a sentence that implies a question such “I wonder if”.
- SVO: a structure that contains a well defined sentence consisting of the subject phrase, verb phrase, and object phrase.
- Block: a block is a fragment of sequential words extracted from a text and produced by preprocessing tasks to enact the parts of speech (either from a sentence, question or noun phrase).
- Synset: a set of synonyms or words related by their meanings.

3.2. Description of the Method

In our research, we pursue a method that describes and processes knowledge in a transparent way by producing results that are always human-readable and can be inspected at any part of the process. Knowledge exploitation of dictionaries or other human-readable knowledge exploitation is key in this process; therefore, we avoid the use of other forms of representations such as word embedding, domain-oriented statistics or probabilities, or tuning of thresholds through training. Besides the use of dictionaries, we use natural language syntax rules for the detection of the three basic forms of expression in any natural language, namely noun phrases, sentences and questions. Our approach to capturing knowledge and establishing the similarity between texts considers a set of syntax rules of a natural language, a model of knowledge processing and a qualitative process for debiasing the knowledge processing. The method is shown in Figure 1; it consists of a perception module that analyses each text to produce one or more blocks of text that may contain sentences, questions or noun phrases. A parsing of each block produces the specification of the lexical types that each word may have; it also retrieves a set of synonyms from a knowledge base and follows the detection of the parts of speech for each block through a set of syntax and disambiguation rules. Once the perception activities are finished, a process of debiasing takes blocks whose content are NPs and tries to integrate them into adjacent sentence clauses; then, sentences and noun phrases are re-analyzed for integration of composed subjects and/or composed objects. The reasoning identifies the similarity between blocks and based on the combination of type of clause, namely sentence, noun phrase, or questions. The last procedure is the assessment consisting of the average of all the similarities found in previous reasoning.

3.3. Perception

The perception module splits a text into blocks according to the punctuation tokens that may indicate SVO clauses, questions or noun phrases. Blocks are parsed using a knowledge base to extract the information on the types of the word the related sets of synonyms (synsets), the possible verbs that exist within the block, the positions of the verbs, the positions of connectives, adverbs, conjunctions and, if possible, the elicitation of the most possible verb according to a set affirmative and destructive rules that re-evaluate and discard possible conjugations of verbs based on the relative position of verb candidates. Finally, a set of syntax rules generates the parts of speech from the verb in each block. The perception module receives as inputs a pair of texts and produces a set of parsed blocks for each text. Table 1 contains the punctuation rules that can be applied to a text; the types of punctuation that a text may have are sentence, question, and conjunctions. The set of rules describes several verbal phrases including present, future, past, continuous forms, perfect tenses, modals and several destructive rules that describe impossible verb phrase scenarios. The outputs are two sets of parsed blocks (one set of parsed blocks per text) and their elicitation of the parts of speech that are used at the debiasing. Algorithm 1 describes

the process for detecting the blocks, parsing the possible type of lexicon of the words and extracting their synsets at each block. A knowledge base is consulted for acquisition of the aforementioned data. Once the types of words are extracted from the KB, an early analysis of possible verb phrases is analyzed.

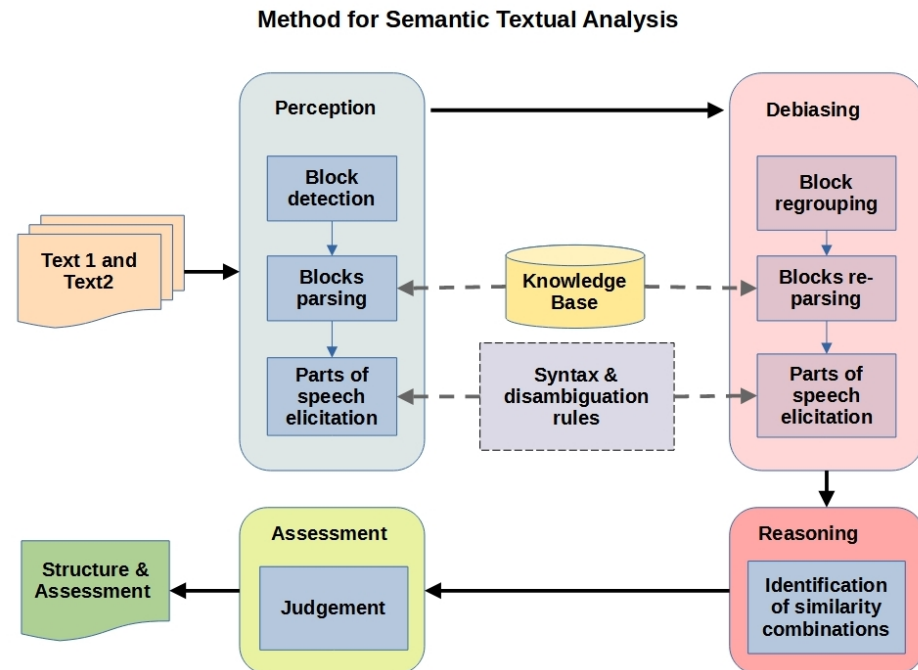


Figure 1. Method for semantic textual similarity analysis.

Table 1. Punctuation and verb phrase rules implemented with perception.

Type	Structures	Blocks
SVO /NP	$(\langle word \rangle)^+ \langle dot \rangle$	1 Block
Q	$(\langle word \rangle)^+ \langle QuestionMark \rangle$	1 Block
Conjunction	$((\langle word \rangle)^+ (\langle comma \rangle, \langle semicolon \rangle))^+ (\langle conjunction \rangle) (\langle word \rangle)^+$	>1 Blocks
Present	present; present continuous; present perfect; conditionals 0, 1, 2 3	>1 Blocks
Past	pas; past continuous; past perfect	>1 Blocks
Future	future; future continuous;	>1 Blocks

Algorithm 1 Perception

Input: Text, KB as knowledge base

Output: A set of Blocks containing SVOs, questions and noun phrases extracted from the text

```

1: PunctuationList ← DetectPunctuation(Text)
2: Blocks ← GetBlocks(PunctuationList)
3: bsize ← size(Blocks)
4: blocktype = getType(PunctuationList, psize)
5: for i = 1 to bsize do
6:   wordsize ← size(Blocks(i))
7:   for j = 1 to wordsize do
8:     Blocks(i).word(j).types ← getTypes(Block(i).word(j), KB)
9:   end for
10: end for
11: Blocks.VerbTenses ← FindVerbtenses(Blocks)
  
```

Perception Example

Consider the perception analysis described in Algorithm 1 on the sentence in Table 2. After the punctuation, signs are pinpointed at the position of the word *as* and the dot sign by applying the rules specified in Table 1. The perception produces two blocks that are parsed separately by extracting the possible types that a word may have in a thesaurus. Finally, the rules for detecting the verb phrases are applied and only one verb phrase is detected on the first block.

Table 2. Perception example.

Sentence	Blocks	Parsing	Verb Phrases
A woman is working as a nurse.	1: A woman is working as	1. A: DETERMINER 2. woman: ADJECTIVE, NOUN 3. is: TO BE 4. working: COUNTABLE, NOUN, PLURAL, 'SINGULAR' UNCOUNT, VERB 5. as: CONJUNCTION	indices: 3, 4
	2: a nurse	1. a: DETERMINER 2. nurse: NOUN	

3.4. Debiasing

Debiasing is the process that removes inaccuracies on the perceived blocks of a text due to the random structure of the latter. The debiasing check for noun phrases can be integrated to SVO clauses or questions. The process for debiasing starts with the identification of noun phrases and the adjacent SVO and questions; if the noun phrases are before a sentence, then the noun phrase becomes part of the sentence and its block is disregarded. On the other hand, if the noun phrase is located after the sentence, then the noun phrase becomes part of the object of the sentence. The exception to this rule of debiasing is when the sentence ends with a period or the noun phrase. In this case, the adjacent has a period at the end; therefore, the noun phrase is not part of the next sentence. Algorithm 2 shows the detailed process with polynomial complexity of $O(n^2)$. The input is the set of blocks obtained from the text and the output is the set of integrated blocks. The first step is the detection of blocks with noun phrases. Next, for each detected noun phrase, the adjacent previous sentence *PreClause* and next sentence *PostClause* are obtained with regards the position that the noun phrase has in the sequence of blocks. If the *PreClause* and *PostClause* are empty, then the noun phrase is added to the *Integrated_Blocks*. If the *PreClause* is non-empty and *PostClause* is empty, then an integration of the noun phrase to the object of the *PreClause* is generated and added to *Integrated_Blocks*. If the *PreClause* is empty and *PostClause* is non-empty, then an integration of the noun phrase to the object of the *PostClause* is generated and added to *Integrated_Blocks*.

Example of the Debiasing Process

Consider the example sentence in Table 2. To remove the inaccuracies that the perception may produce, the blocks are analyzed to integrate parts that may be segregated from SVO clauses. Algorithm 2 searches for blocks with no SVO (lack of verb phrases) to add them to the previous or next block as part of the object or subject, respectively. In the considered example, only two NPs appear in the text: “a, woman” and “a nurse”. Furthermore, only the first block has a verb phrase. Therefore, the noun phrase “a woman” has a preclause; however, the next block has only an NP, so it follows that the second condition is applied and the content of the second block is merged with the first block as part of the object of the SVO contained in the first block. Table 3 shows the result of applying the debiasing to obtain a single SVO by the integration of two blocks produced by the perception process.

Algorithm 2 Debiasing**Input:** a set of Blocks containing SVOs, questions and noun phrases extracted from the text**Output:** Integrated_Blocks extracted from Blocks

```

1:  $DetectedNPs \leftarrow ExtractNPs(Blocks)$ 
2:  $npsize \leftarrow size(DetectedNPs)$ 
3:  $Integrated\_Blocks \leftarrow \emptyset$ 
4: for  $i = 1$  to  $npsize$  do
5:    $PreClause \leftarrow ObtainPreAdjacentClause(Blocks, DetectedNPs(i))$ 
6:    $PostClause \leftarrow ObtainPostAdjacentClause(Blocks, DetectedNPs(i))$ 
7:   if  $PreClause = \emptyset \wedge PostClause = \emptyset$  then
8:      $Integrated\_Blocks \leftarrow add(Integrated\_Blocks, Detected(NPs(i)))$ 
9:   end if
10:  if  $PreClause \neq \emptyset \wedge PostClause = \emptyset$  then
11:     $NewClause \leftarrow IntegrateObject(PreClause, DetectedNPs(i))$ 
12:     $Integrated\_Blocks \leftarrow add(Integrated\_Blocks, NewClause)$ 
13:  end if
14:  if  $PreClause = \emptyset \wedge PostClause \neq \emptyset$  then
15:     $NewClause \leftarrow IntegrateSubject(PreClause, DetectedNPs(i))$ 
16:     $Integrated\_Blocks \leftarrow add(Integrated\_Blocks, NewClause)$ 
17:  end if
18: end for

```

Table 3. Debiasing example.

Sentence	Blocks	Parsing	Verb Phrases
A woman is working as a nurse.	1: A woman is working as a nurse.	1. A: DETERMINER 2. woman: ADJECTIVE, NOUN 3. is: TO BE 4. working: COUNTABLE, NOUN, PLURAL, 'SINGULAR' UNCOUNT, VERB 5. as: CONJUNCTION 6. a: DETERMINER 7. nurse: NOUN	indices: 3, 4

3.5. Reasoning

Reasoning is the part where blocks from the two different texts are aligned by two criteria: the raw similarity and the speech similarity. The raw similarity is detected by the use of matching terms, the synsets of the terms and the morphology of pairs of terms. Figure 2 shows a series of text snippet combinations that the system can handle. The mathematical models for dealing with the cases of combinations developed in this subsection are described in Section 3.5.1. The details of the combinations of texts are described in Section 3.5.2.

3.5.1. Semantic Similarity Judgement

The assessment of semantic text similarities considers the scenarios where the pairs of text may involve one of six text combinations of noun phrases, sentences and questions. This section introduces the primary equations for assessing the similarity based on the appropriate text combination and their content. The definitions stated in this section are exploited according to each combination case developed in Section 3.5.2.

Definition 1 (Raw similarity). Let a pair of texts T_x and T_y , raw the similarity of a pair of texts is given as the degree of common knowledge or Jaccard similarity specified by the matching of their purged sets of stopwords (A stopword is a word that is widely used in texts regardless the domain of text, as example determiners, pronouns, conjunctions, adverbs, as well as a few verbs as be, have, do.), and S_1 and S_2 the set of words (1) from T_x and T_y respectively.

$$RawSim(S_1, S_2) = \frac{\|S_1 \cap S_2\|}{\|S_1 \cup S_2\|} \quad (1)$$

The Raw Similarity produces values in $[0, 1]$, this similarity is used as a first detection of similarity between two texts. This early similarity is also used for assessment only in the case of comparison between two noun phrases and the extraction of similarities of subjects and objects of sentences.

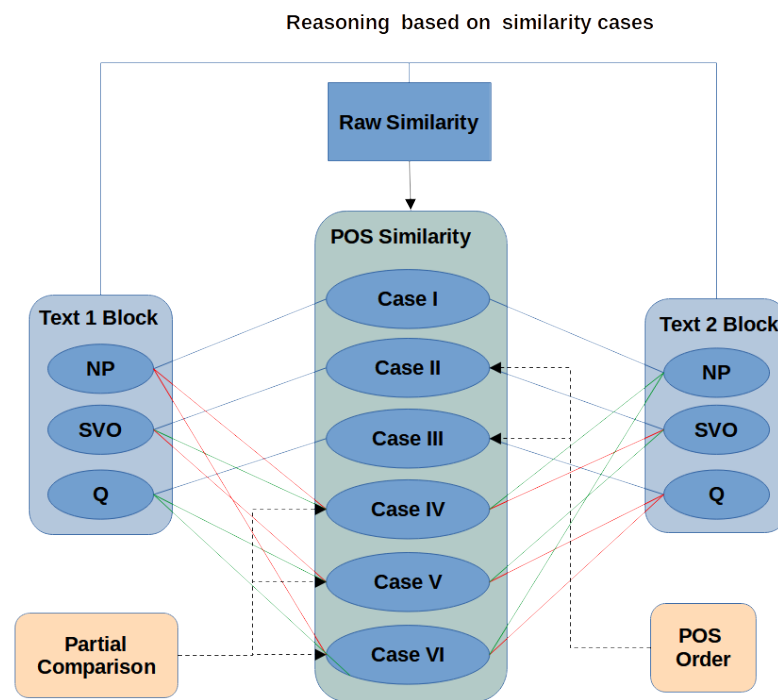


Figure 2. Reasoning map for similarities detection.

Definition 2 (Similarity of synonyms). Let a pair of texts T_x and T_y , the similarity by synonymy of a pair of words is given as the degree of common knowledge specified by the matching (2) of the synonym sets (synsets) where T_x and T_y are purged of stopwords which are extracted from a knowledge base, dictionary or thesaurus.

$$M(S_1, S_2) = \begin{cases} 1 : S_1 \cap S_2 \neq \emptyset \\ 0 : S_1 \cap S_2 = \emptyset \end{cases} \quad (2)$$

The similarity of synonyms is used for the detection of semantics on differentiated word forms. This type of similarity exploits a knowledge base to get the sets of synonyms (synsets). The similarity of synonyms is used to improve the assessment of similarity when synonyms are found in a pair of text snippets.

Definition 3 (Similarity of noun phrases). Let a pair of texts T_x and T_y containing only NPs, the similarity of a pair of NPs is defined as the degree of common knowledge specified by the matching (3) of the synonym sets $T_x(i).S$ and $T_y(k).S$ of a pair of words in T_x and T_y , where T_x and T_y are purged of stopwords which are extracted from a knowledge base, dictionary or thesaurus.

$$S(T_x, T_y) = \frac{\sum_{i=1}^{|T_x|} \sum_{k=1}^{|T_y|} M(T_x(i).S, T_y(k).S)}{|T_x \cup T_y|} \quad (3)$$

The similarity of noun phrases is used on the combination of texts containing only noun phrases or subjects and objects from sentences and questions.

Definition 4 (Similarity of sentences or questions). Let a pair of texts T_x and T_y containing SVOs (sentences) and/or VSOs (questions), the similarity of a pair of SVO and/or VSO is given as the degree of common knowledge specified by the matching of the parts of speech at the subjects $S1 \in T_x$ and $S2 \in T_y$, verbs $V1 \in T_x$ and $V2 \in T_y$, and objects $O1 \in T_x$ and $O2 \in T_y$, weighed by a precedence factors for the subject $subf$ and the verb $verbf$ (4).

$$Sim(T_x, T_y) = \frac{S(S1, S2) * subf + S(V1, V2) * verbf + S(O1, O2)}{3} \quad (4)$$

The similarity of sentences or questions is applied case by case to the combination of sentences and/or questions found in a pair of text snippets. If the verbs in the sentences have no semantic similarity, then an average of the similarity of the subject and object is computed with an $\alpha \in [2, 2.5]$ as described in Equation (5); the α constant has a partial discount due to the absence of similarity for verbs.

$$SimNV(T_x, T_y) = \frac{S(S1, S2) + S(O1, O2)}{\alpha} \quad (5)$$

3.5.2. Cases of Combinations on POS Alignment

The combinations of text snippets are defined in Table 4; these cases require specific criteria based on the structure and content of each text snippet. Two assessments are defined for each case—the analysis of parts of speech and the integration analysis. The part of speech analysis assesses the content at the different parts of speech. The integration analysis may include a penalty criterion due to the structural differences. Cases I, II, and III share similar structures; thus, the criteria are applied straightforwardly without discount factors. Cases IV and VI have discount factors after the assessment due the lack of a verb in one of the snippets. Case V applies a discount factor due to one of the texts containing a sentence and the other containing a question.

Table 4. Combinations of pair types.

Case	Type	Assessment Criteria	POS	Integration
I	NP_x NP_y	— The noun phrases are compared on morphology, use of synonyms and by direct matching	(1), (2)	$S(NP_x, NP_y)$ (3)
II	SVO_x SVO_y	— The sentences are compared on the parts of speech (subject, object, and verbs), an order analysis is applied to identify if the main actors are related.	(3)	$Sim(SVO_x, SVO_y)$ (4) $SimNV(SVO_x, SVO_y)$ (5)
III	$Q_x - Q_y$	The questions are compared on the parts of speech (subject, object, and verbs), an order analysis is applied to identify if the main actors are related.	(3)	$Sim(Q_x, Q_y)$ (4)
IV	SVO NP	— The sentence is compared using its subject and object against the noun phrase. A discount factor is applied due to the absence of a verb.	(1), (2)	$0.75 * S(SVO.S \cup SVO.O, NP_y)$ (3)
V	$SVO - Q$	The sentence and the question are compared at their subjects, objects and verbs, a discount factor is applied due to the interrogative nature of the latter.	(3)	$0.75 * Sim(SVO, Q)$ (4)
VI	$Q - NP$	The sentence is compared using its subject and object against the noun phrase. A discount factor is applied due to the absence of a verb.	(1), (2)	$0.75 * S(Q.S \cup Q.O, NP_y)$ (3)

Similarity by parts of speech uses raw similarity of subjects, verbs, and objects and the ordering of coincidences of the raw similarity of the parts of speech. For a better understanding, consider Figure 2 which describes the cases for similarity detection that the machine applies to mimic the reasoning of similarity from scratch. The reasoning starts with a detection (1) of a pair of blocks regardless of the type of structure that they contain

(SVO, questions, or a noun phrase). If the detected raw similarity is non-zero, a part of speech similarity selects one criterion from a set of six cases. Case I establishes the similarity between a pair of noun phrases as an NP similarity defined in (3). Case II takes as inputs a pair of SVO blocks; in this case, an analysis of similarities is applied to the pairs of subjects, objects and verbs (4). It includes order analysis of the parts of speech consisting of testing whether a concepts appears in the same part of speech at both SVOs. Case III analyses pairs of question blocks at the level of their parts of speech (verbs, subjects, and objects); this processing is analogous to the extraction of SVO similarities in Case II (4). The order analysis is also applied to the parts of speech. Case IV finds the similarity between an SVO and an NP (3); in this case, the subjects and objects of the SVO are compared to the NP and the verb of the SVO is considered as a difference. Case V obtains the similarity of the parts of speech for an SVO and a question Q (4); in this case, the subjects, objects, and verbs are compared and order analysis is applied on the subjects and objects. Case VI analyses the similarity between a question and a noun phrase (3); this case is analogous to Case IV due to the subject and object of the question being compared to the noun phrase and the existence of the verb in the question being considered a difference. Algorithm 3 shows the implementation steps to perform the reasoning depicted in Figure 2; it starts with the extraction of the RawSim from a pair of blocks extracted from different texts. If the RawSim is non-zero, then based on the types of both blocks (SVO, question, or noun phrase), only one of the six criteria is applied to detect the degree of similarity and the differences. The similarities produced by the reasoning are the input of the assessment.

3.5.3. Examples of Reasoning on Combinations

As example of the combinations of types on pairs of texts, consider the samples extracted from the SemEval test and training datasets [33] shown in Table 5. The computation of the RawSim (1) is shown in Table 6. Based on the signs of punctuation at Table 1, one case is applied to a pair of blocks extracted from the texts. The results of the RawSim at each pair lead to a further analysis on the assessment defined in the next subsection.

Table 5. Example of reasoning on a pair of texts from the SemEval 2017 dataset.

Case	Text 1	Text 2
I	A young person deep in thought.	A young man deep in thought.
II	A person is on a baseball team.	A person is playing basketball on a team.
III	How exactly is Germany being ‘punished’ for the stupidity of WW?	How exactly are they being punished?
IV	A dog under the stairs	A dog is resting on the stairs.
V	We never got out of it in the first place!	Where does the money come from in the first place?
VI	Why are Russians in Damascus?	Russians in Damascus!

Table 6. Early similarity using Jaccard of pairs on Table 5.

Case	Computation (1)	Sim > 0
I	$RawSim(NP_1, NP_2) = \frac{\ \{young, deep, thought\}\ }{\ \{young, person, deepman, thought\}\ } = \frac{3}{5}$	yes
II	$RawSim(SVO_1, SVO_2) = \frac{\ \{person, team\}\ }{\ \{baseball, basketball, person, team\}\ } = \frac{2}{4}$	yes
III	$RawSim(Q_1, Q_2) = \frac{\ \{germanystupidity\}\ }{\ \{Germany, stupidity, WW\}\ } = \frac{2}{3}$	yes
IV	$RawSim(SVO_2, NP_1) = \frac{\ \{dog, stairs\}\ }{\ \{dog, stairs\}\ } = \frac{2}{2}$	yes
V	$RawSim(SVO_1, Q_2) = \frac{\ \{place\}\ }{\ \{place, money, place\}\ } = \frac{1}{3}$	yes
VI	$RawSim(Q_1, NP_2) = \frac{\ \{russians, damascus\}\ }{\ \{russians, damascus\}\ } = \frac{1}{1}$	yes

Algorithm 3 Similarity reasoning

Input: Two sets of blocks Blocks1 and Blocks2 containing SVOs, questions and noun phrases extracted from the text

Output: SimSet as a set of similarities extracted from the Blocks of both

```

1: block1size ← size(Blocks1)
2: block2size ← size(Blocks2)
3: Similarities ← ∅
4: simcounter ← 0
5: for i = 1 block1size do
6:   for j = 1 block2size do
7:     RawSim ← RawSim(Blocks1(i), Blocks2(j))
8:     if RawSim > 0 then
9:       if Blocks1(i).Type = NP ∧ Blocks2(j).Type = NP then
10:        simcounter ← simcounter + 1
11:        SimiSet(simcounter).similarity ← CaseI(Blocks1(i), Blocks2(j))
12:       end if
13:       if Blocks1(i).Type = SVO ∧ Blocks2(j).Type = SVO then
14:        simcounter ← simcounter + 1
15:        SimiSet(simcounter).similarity ← CaseII(Blocks1(i), Blocks2(j))
16:       end if
17:       if Blocks1(i).Type = Question ∧ Blocks2(j).Type = Question then
18:        simcounter ← simcounter + 1
19:        SimiSet(simcounter).similarity ← CaseIII(Blocks1(i), Blocks2(j))
20:       end if
21:       if Blocks1(i).Type = SVO ∧ Blocks2(j).Type = NP then
22:        simcounter ← simcounter + 1
23:        SimiSet(simcounter).similarity ← CaseIV(Blocks1(i), Blocks2(j))
24:       end if
25:       if Blocks1(i).Type = SVO ∧ Blocks2(j).Type = Question then
26:        simcounter ← simcounter + 1
27:        SimiSet(simcounter).similarity ← CaseV(Blocks1(i), Blocks2(j))
28:       end if
29:       if Blocks1(i).Type = Question ∧ Blocks2(j).Type = NP then
30:        simcounter ← simcounter + 1
31:        SimiSet(simcounter).similarity ← CaseVI(Blocks1(i), Blocks2(j))
32:       end if
33:     end if
34:   end for
35: end for

```

3.6. Assessment

The assessment is obtained according to the findings based on the six classes of similarities of reasoning. The process of generating a judgement produces an average of the similarities and multiplies the average by five. The similarity judgement is specified according to the criteria defined in Table 4; based on each case, additional criteria are applied to address specific differences, and such differences imply reduced similarities in the interval of $[0, 5]$ to comply with the gold standard defined by [33].

Examples of the Assessment Process

Table 7 describes the assessment of the combinations developed in the reasoning. All the assessments are computed first in the interval of $[0, 1]$, then the assessment is recomputed in the scale of $[0, 5]$ according to the SemEval 2017 contest on Task 1 track5 (en-en) [33]. The example in Case I has an assessment of 5 due to person and man belonging to the synset and the rest of words presenting both sets as the same. The example in Case II applies four computations—three for the alignment of subjects, verbs and objects, respectively, and a fourth for the assessment of the integration of previous POS computations. Based on the content of the pair, there is similarity between the subjects, no similarity between the verbs, and full similarity between the synsets of *basketball* and *baseball*. Since the verbs are different, the equation that considers a discount factor is chosen (5), and the factor α is set to 2.5 since there exists at least one similarity not deducted by exact match. The example in Case III has three previous computations at the POS of the questions regarding the verbs, subject, and objects, respectively. The integration of the similarity is the average obtained from the computations on the POS. The content of the pair has only a non-zero similarity at the objects of the questions; therefore, the similarity is set to 1.65. The example in Case IV has only one integration of similarity for the union of the subject and object of the SVO and the NP. In the content of the example, the similarity is the highest; however, a discount factor set the integration at 3.75 instead. For Case V, three computations on the POS are computed on the subjects, verbs, and objects of the SVO and the question. Due to their differentiated nature, a similarity discount factor reduces the POS similarities to 75% of the average. From the content of the pair, only the respective objects have a non-zero similarity. Additionally, the discount factor is applied; therefore, the similarity is set to 1.24. The example in Case VI has only an integration of similarity regarding the union of the subject and object of the question against the NP. A discount factor is applied to the computation of the result. The similarity of the content achieves the highest similarity; however, the discount factor sets the similarity to 3.75.

Table 7. Similarity assessment of pairs from Table 5.

Case	Integration	Sim
I	(3) $S(NP_1, NP_2) = \frac{M(young, young) + M(deep, deep) + M(thought, thought) + M(person, man) + M(man, person)}{\ \{young, person, deep, man, thought\}\ } = \frac{5}{5} = 1$	5.0
II	$(3) S(Sub_1, Sub_2) = \frac{M(person, person)}{\ \{person\}\ } = 1$ $(3) S(V_1, V_2) = \frac{0}{\ \{be, work\}\ } = 0.0$ $(3) S(Obj_1, Obj_2) = \frac{M(basketball, baseball) + M(baseball, basketball) + M(team, team)}{\ \{team, basketball, baseball\}\ } = 1$ $(5) SimNV(SVO_1, SVO_2) = \frac{S(Sub_1, Sub_2) + S(V_1, V_2) + S(Obj_1, Obj_2)}{\alpha} = \frac{1 + 0 + 1}{2.5} = 0.8$	4.0
III	$(3) S(Sub_1, Sub_2) = \frac{0}{\ \{germany, they\}\ } = 0$ $(3) S(V_1, V_2) = \frac{M(punish, punish)}{\ \{punish\}\ } = 1$ $(3) S(Obj_1, Obj_2) = \frac{0}{\ \{ww, stupidity\}\ } = 0$ $(4) Sim(Q_1, Q_2) = \frac{S(Sub_1, Sub_2) + S(V_1, V_2) + S(Obj_1, Obj_2)}{3} = \frac{0 + 1 + 0}{3} = 0.33$	1.65
IV	(3) $S(Sub_1 \cup Obj_1, NP_2) * 0.75 = \frac{M(dog, dog) + M(stairs, stairs)}{\ \{dog, stairs\}\ } \times 0.75 = 0.75$	3.75
V	$(3) S(Sub_1, Sub_2) = \frac{0}{\ \{we, money\}\ } = 0$ $(3) S(V_1, V_2) = \frac{0}{\ \{getout, come\}\ } = 0$ $(3) 3 S(Obj_1, Obj_2) = \frac{M(first, first) + M(place, place)}{\ \{first, place\}\ } = 1$ $(4) 0.75 * Sim(SVO_1, Q_2) = \frac{S(Sub_1, Sub_2) + S(V_1, V_2) + S(Obj_1, Obj_2)}{3} = 0.75 \times \frac{0 + 0 + 1}{3} = 0.24$	1.24
VI	(3) $S(Sub_1 \cup Obj_1, NP_2) \times 0.75 = \frac{M(russians, russians) + M(damascus, damascus)}{\ \{russians, damascus\}\ } \times 0.75 = 0.75$	3.75

4. Experiments

An experiment is performed to test the accuracy of the method on semantic similarity. The objective of the experiment is to provide an insight on the performance according to the number of concepts and their synonym relationships on several knowledge bases.

Experiment 1. The method is tested on seven knowledge bases numerated from KB1 to KB7 and conducted on a combination of six dictionaries and two thesauruses that contain different amounts of concepts and synsets, respectively. Experimental data are the “test” dataset and the gold standard from SemEval 2017 [33]. The “test” dataset consists of 250 text snippet pairs in English. The cumulative set of words within the dataset has 872 words after the removal of frequent use words (stopwords). The assessment scale used by the SemEval 2017 Task 1 as described in [33] is [0, 5], where 0 means no similarity and 5 means full similarity. The combinations of knowledge bases are described in Table 8. For each combination, a Pearson correlation is computed from the comparison of each result and the gold standard. The results are depicted using a chart with a range of [−5, 5] for the identification of underestimation (values in [−5, 0]) and overestimation (values in [0, 5]), where each point in the chart identifies the difference of the assessed similarity of a pair using the method on a given knowledge base and the gold standard defined by SemEval 2017 for such a pair.

Table 8. Knowledge bases tested on experiment 1.

KB	Dictionary	Thesaurus	Found Concepts	Missing Concept
1	Collins	Thesaurus.com	826/802	46/70
2	Wordnet	Synonym.com	800/827	72/45
3	Dictionary.com	Synonym.com	828/827	44/45
4	MacMillan	Synonym.com	621/827	251/45
5	Oxford	Synonym.com	733/827	139/45
6	Merriam Webster	Synonym.com	798/827	74/45
7	Collins	Synonym.com	826/827	46/45

5. Results

The results of the seven combinations are depicted in the charts contained in Figure 3, where a value below zero means a subestimation of our results and a value above zero means an overestimation of our results. If the assessed similarity and the gold standard match, the difference is zero. The Pearson correlations achieved for the used knowledge bases are shown in Table 9. The results show that KB1 has the best Pearson correlation and has the second-most complete dictionary and the least complete thesaurus in “Thesaurus.com”. The most complete in terms of numbers of concepts is KB3; however, it has a lower performance than other knowledge bases that are more incomplete due to the quantity of relations of synonymy that a concept contains in the knowledge base. Figure 3 shows the difference between our results and the SemEval 2017 [33] gold standard by exploiting the seven knowledge bases described in Table 8. In these charts, the objective is that the difference between the results of the method and the gold standard is zero. A positive value is an overestimation of our method (a degree of positive falsehood) and a negative value is a sub-estimation (a degree of negative falsehood); the results of each knowledge base are consistent with the Pearson correlation achieved by the latter. Our results demonstrate that the method takes advantage of the content when the knowledge is more complete regardless of the classification of the knowledge. It is not surprising that a knowledge base with less concepts achieves better results than another knowledge base that has more concepts. The reason for this is the number of synonymy relations of the concepts in a knowledge base; a concept with more relations of synonymy has more opportunities to contribute to the assessment. The number of synonym relations contained in the concepts from Synonym.com is lower than the number of synonym relations that concepts from Thesaurus.com have. The differences between the proposed method and the rest of methods enlisted in this research are the following: (1) our method uses unlabeled data, whereas the other methods exploit other systems to label the data prior the preprocessing of datasets; (2) our method may consider multiple semantics and the knowledge is represented in a human-readable way; (3) our method is independent of the content and the classification is of linguistic types (noun phrases, sentences, and question) instead of domains of knowledge. Compared against supervised methodologies [9–11], our method has the advantage of no need for assumptions of the knowledge, as the lack of assumptions is mitigated by the content in the knowledge base and syntax-driven processing, and the biases of the knowledge base are handled with the extraction of synonyms of related concepts within the knowledge base. One disadvantage of our method is that in the generalized way of assessment, the amount of information of the knowledge base and small biases of the perception and the reasoning produce minor accuracy compared to the supervised methods. The processing of text independent from the knowledge domain is important for the creation of adaptable systems in uncertain scenarios; as an example, we compare our method with the top methods considered in [33] regarding several of the most challenging pairs in the test dataset in the following subsection.

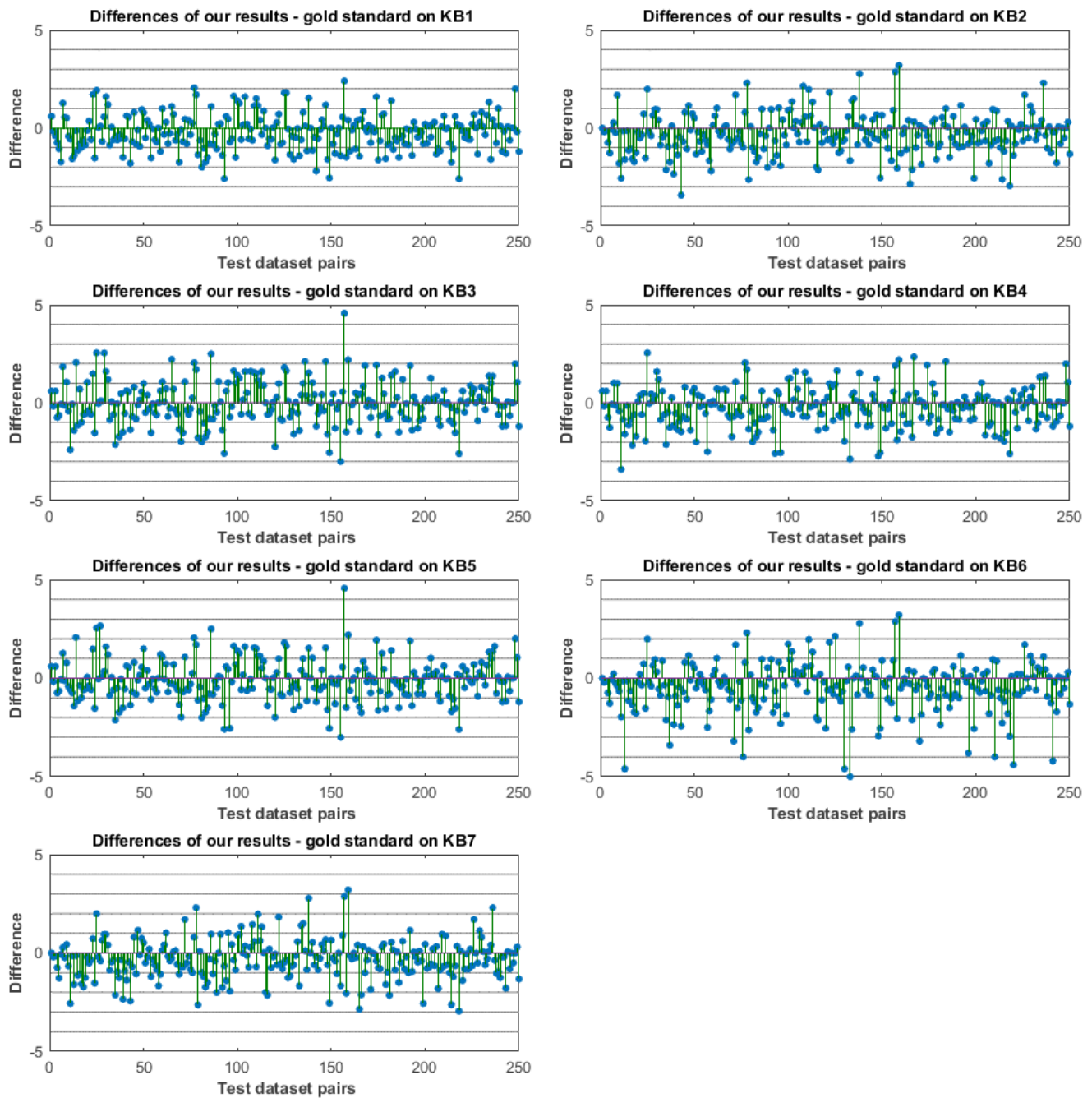


Figure 3. Results of the semantic text similarity of the proposed method on 7 knowledge bases.

Table 9. Pearson correlations on the tested knowledge bases.

KB 1	KB 2	KB 3	KB 4	KB 5	KB 6	KB 7
77.46%	70.43%	68.12%	71.63%	70.40%	57.16%	72.44%

5.1. Comparison of the Experiment with Related Algorithms

For demonstrative purposes, we show the performance of the most relevant methods on the most difficult pairs of sentences in the test dataset used in SemEval [33]. Table 10 includes six of the most difficult pairs of sentences tested on the top models [9–11,31,34], the gold standard of the pairs of sentences, and the results of the proposed method; the closest score to the gold standard for each pair is highlighted in blue. Our method is closer to the gold standard (GS) in four of the pairs due to the the analysis of the parts of speech performed by the proposed method.

Table 10. Comparison of the assessment between the top systems reported by [33], its gold standard and our proposed method, on the six most difficult pairs in the Test dataset.

SemEval 2017 Pairs			Top Models		Proposed Method	
Pairs [33]	(GS) [33]	Score	(Difference)	Model	Our	(Difference)
Pair 14	1.8	3.2	(+1.4)	DT_team [9]	1.67	(−0.13)
Pair 78	1.0	1.9	(+0.9)	FCICU [31]	2.70	(+1.70)
Pair 84	4.0	3.6	(−0.4)	BIT [11]	2.50	(−1.50)
Pair 115	5.0	4.5	(−0.5)	ITNLP [34]	5.00	(+0.00)
Pair 184	3.0	4.0	(+1.4)	BIT [11]	2.50	(−0.50)
Pair 195	0.2	0.8	(+0.6)	FCICU [31]	0.14	(−0.06)

5.2. Performance of the Proposed Method with Regards the State of the Art

Due to our method not including training, and the majority of the methods requiring training or use of resources produced by systems that learned the knowledge through training, the proposed method has more in common with some state-of-the-art unsupervised models. Despite the differences between the analytical approach of our method and the training approach adopted by most models, in this subsection, our method is compared with the results obtained by the state-of-the-art models. The comparison includes the models from SemEval 2017 Task 1 (Semantic Text Similarity) and latter semantic text similarity models using BERT models. Table 11 contains the results of Pearson correlation on the SemEval test dataset, the models' state-of-the-art features, and the proposed method. Table 11 shows the models and their Pearson correlation results reported in [33], if models use synsets, the alignment of concepts or sentences (Align), training (Train), knowledge bases (KB), word embeddings (WEmb), management of stopwords, parts of speech (POS) and the type of assessment of similarity (Sim). Our proposed method has the rare feature of analyzing the POS at sentence level (SVO), and also does not require resources that have been enriched through learning (in contrast to FCICU which uses the StanfordCoreNLP and Wordnet of BabelNet, or BIT which uses statistical frequencies and NLTK toolkits; these toolkits contain resources generated through the training of models). The majority of the models require embeddings which are refined through learning. On the other hand, only three models exploit synsets (including the proposed model). From the results in Table 8, we can observe that the content of each knowledge base has a direct impact in Table 9; in this regard, our proposal has the advantage of adding new concepts (in a human-readable way) to a knowledge base is sufficient to operate with updated knowledge.

Table 11. Baseline of models used on SemEval 2017 (using Pearson correlation $\times 100$) divided by supervised and unsupervised models.

Model	Pearson $\times 100$	Synsets	Align.	Train	KB	WEmb	StopWords	POS	Sim
Supervised									
ECNU [10]	85.18	-	BOW, Dependency,	DAN [18], LSTM [8]	PPDB [35]	Glove [36], Paragram	-	Stanford [19]	Regression (RF, GB)
BIT [11]	84.00	-	-	LR, SVM	British National Corpus	Word2vec [13] + IDF	-	NLTK [37]	cosine + IDF
DT_team [9]	83.60	-	Word and chunk	DSSM, CDSSM	PPDB [35]	Word2vec [13], Sent2vec	-	own POS	LR, GB
ITNLPaiKF [34]	82.31	-	Semantics, context	SVR	Wikipedia, twitter	Word2vec [13], Glove [36]	-	NLTK [37]	stat. freq. (IC) [20]
MITRE [38]	80.53	-	based on cosine	CRNN, LSTM	Wikipedia	Word2vec [13]	-	-	string sim.
HCTI [39]	81.56	-	-	CNN	GloVe [36]	GloVe [36]	-	NLTK [37]	Cosine
Udl [40]	80.04	-	Align. POS	Reg. RF	GloVe [36]	GloVe [36]	-	-	cosine
STS-UHH [41]	80	-	Glove [36], Dependency Graph	LDA	Distributional Thesaurus	Glove [36]	-	Stanford [19], TLCS	weighted cosine
PurdueNLP [42]	79.28	-	-	Skip-Gram	PPDB [35]	Paraphrase and Event	-	-	Regression
neobility [43]	79.25	-	N-gram overlap	RNN, GRU	Wikipedia, Wordnet [7]	Word2vec [13]	-	-	Cosine
OPI-JSA [44]	78.50	-	-	RNN, MLP	BNC, Book-Corpus	GloVe [36]	-	PoS weighted on cosine	cosine
L2F/INESC-ID [45]	78.11	-	-	NN	SICK	Vectors	-	-	SMATCH
Lump [46]	73.76	from Babel-Net [5]	Explicit Analysis	GB, SVM	BabelNet [5]	Word2vec [13]	-	-	Cosine
ResSim [47]	69.06	-	Word Align.	NN	Europarl	Europarl	-	-	Adam alg.
Unsupervised									
FCICU [31]	82.80	from Babel-Net [5]	Similarity metric	-	BabelNet [5]	-	yes	Stanford [19]	Synset, Alignment
BIT [11]	81.61	-	Sentence Alignment	-	British National Corpus	-	-	NLTK [37]	stat. freq. (IC) [20]
SEF@UHH [32]	78.80	-	-	PV-DBOW	Common-crawl, others	Doc2Vec [48]	-	-	cosine
Our	77.40	deducted from KB	Driven by POS	-	Dictionary, Thesaurus	-	yes	Rules (SVO)	Weighted by POS
STS-UHH [41]	73	-	GloVe [36]	-	-	Glove [36]	-	Stanford [19]	Cosine
QLUT [49]	68.87	-	-	-	Wikipedia	Word2vec [13]	yes	Stanford [19]	Cosine

6. Conclusions

This work presents a method for semantic text similarity that considers perception, debiasing, and reasoning on similarities for the assessment of similarities from scratch and exploiting a database. A set of polynomial algorithms exploit stored information in knowledge bases and decompose the texts into blocks to facilitate the identification of similarities throughout the parts of speech of a pair of texts. A process of debiasing corrects the inaccuracies of the identification of the structure of the texts. A process of reasoning

classifies the texts according to their syntactic properties. A process of assessment computes the average similarity produced by the comparison of blocks elicited from a pair of texts. The method is implemented on a system that uses the “test” dataset of SemEval 2017 and seven knowledge bases integrated with seven dictionaries and two thesauruses. The results demonstrate that the system extracts information without any more assumptions than the grammar of English; the method is resilient to biases and its performance relies on the degree of completeness of the knowledge base in use. The method enables optimization of computational resources since it requires no training or resources produced by training. The integration of our methodology into fully domain-related knowledge representations such Wordnets for improvement of the assessment is an avenue for future work.

Author Contributions: Conceptualization, O.Z.; methodology, J.Y.R.-M.; software, O.Z. and S.R.-C.; validation, G.O.-T. and F.d.J.S.-V.; data curation, I.G.-E. and J.C.M.-S.; writing—original draft preparation, O.Z. and S.R.-C.; writing—review and editing, F.d.J.S.-V. and G.O.-T.; funding acquisition, J.Y.R.-M., I.G.-E. and J.C.M.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The code for reproducing the experiments can be found at the following permanent link <https://figshare.com/account/articles/20814406>. The code also contains a version of the Test dataset and its gold standard from SemEval 2017 [33], and its gold standard. The knowledge bases used in this study are available on request from the corresponding author. The knowledge bases are not publicly available due to copyright since it is composed of definitions retrieved from online dictionaries as described in Table 8.

Acknowledgments: The authors thank the anonymous reviewers for their comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Doolittle, P.E. Understanding Cooperative Learning through Vygotsky’s Zone of Proximal Development. In Proceedings of the Lilly National Conference on Excellence in College Teaching, Columbia, SC, USA, 2–4 June 1995. Available online: <https://files.eric.ed.gov/fulltext/ED384575.pdf> (accessed on 17 April 2023).
2. Delprato, D.J.; Midgley, B.D. Some fundamentals of BF Skinner’s behaviorism. *Am. Psychol.* **1992**, *47*, 1507. [CrossRef]
3. Wang, Y.; Zatarain, O.A. A Novel Machine Learning Algorithm for Cognitive Concept Elicitation by Cognitive Robots. *Int. J. Cogn. Inform. Nat. Intell.* **2017**, *11*, 31–46. [CrossRef]
4. Wang, Y. Concept Algebra: A Denotational Mathematics for Formal Knowledge Representation and Cognitive Robot Learning. *J. Adv. Math. Appl.* **2015**, *4*, 61–86. [CrossRef]
5. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [CrossRef]
6. Wang, Y.; Zatarain, O.A. Design and Implementation of a Knowledge Base for Machine Knowledge Learning. In Proceedings of the IEEE 17th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC, Berkeley, CA, USA, 16–18 July 2018; pp. 70–77.
7. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
8. Hochreiter, S.; Schmidhuber, J. Long Short Term Computation. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
9. Maharjan, N.; Banjade, R.; Gautam, D.; Tamang, L.J.; Rus, V. DT_Team at SemEval-2017 Task 1: Semantic Similarity Using Alignments, Sentence-Level Embeddings and Gaussian Mixture Model Output. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 120–124.
10. Tian, J.; Zhou, Z.; Lan, M.; Wu, Y. ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 191–197.
11. Wu, H.; Huang, H.; Jian, P.; Guo, Y.; Su, C. BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 77–84.
12. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
13. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Jeffrey, D. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the NIPS’13: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2016; pp. 3111–3119.

14. Maharjan, N.; Banjade, R.; Rus, V. Automated Assessment of Open-ended Student Answers in Tutorial Dialogues Automated Assessment of Open-ended Student Answers in Tutorial Dia- logues Using Gaussian Mixture Models. In Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, Marco Island, FL, USA, 22–24 May 2017; pp. 98–103.
15. Sarić, F.; Glavaš, G.; Karan, M.; Snajder, J.; Dalbelo, B.; Bašić, B. TakeLab: Systems for Measuring Semantic Text Similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics, Montreal, QC, Canada, 7–8 June 2012; pp. 441–448.
16. Moschitti, A. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Proceedings of the 17th European Conference on Machine Learning Machine Learning: ECML 2006, Berlin, Germany, 18–22 September 2006; Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Eds.; Springer: Berlin/Heidelberg, Germany; pp. 318–329.
17. Sultan, M.A.; Bethard, S.; Sumner, T. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; pp. 148–153. [\[CrossRef\]](#)
18. Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; Iii, H.D. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1681–1691.
19. Manning, C.D.; Bauer, J.; Finkel, J.; Bethard, S.J. The Stanford CoreNLP Natural Language Processing Toolkit Christopher. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014; pp. 55–60.
20. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the IJCAI'95: 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 7, pp. 448–453.
21. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [\[CrossRef\]](#)
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
24. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for self-supervised learning of language representations. In Proceedings of the Eighth International Conference on Learning Representations ICLR 2020, Online, 26 April–1 May 2020.
25. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
26. Xu, C.; Zhou, W.; Ge, T.; Wei, F.; Zhou, M. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7859–7869.
27. Sheng, T.; Wang, L.; He, Z.; Sun, M.; Jiang, G. An Unsupervised Sentence Embedding Method by Maximizing the Mutual Information of Augmented Text Representations. In Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2022, Bristol, UK, 6–7 September 2022; Springer: Cham, Switzerland, 2022; pp. 174–185.
28. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 4163–4174. [\[CrossRef\]](#)
29. Izsak, P.; Berchansky, M.; Levy, O. How to Train BERT with an Academic Budget. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 10644–10652. [\[CrossRef\]](#)
30. Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Zhao, T. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–8 July 2020; pp. 2177–2190. [\[CrossRef\]](#)
31. Hassan, B.; Abdelrahman, S.E.; Bahgat, R.; Farag, I. UESTS: An Unsupervised Ensemble Semantic Textual Similarity Method. *IEEE Access* **2019**, *7*, 85462–85482. [\[CrossRef\]](#)
32. Duma, M.S.; Menzel, W. SEF@UHH at SemEval-2017 Task 1: Unsupervised Knowledge-Free Semantic Textual Similarity via Paragraph Vector. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 170–174.
33. Cer, D.; Diab, M.; Agirre, E.; Iñigo, L.G.; Specia, L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 1–14.
34. Liu, W.; Sun, C.; Lin, L.; Liu, B. ITNLP-AiKF at SemEval-2017 Task 1: Rich Features Based SVR for Semantic Textual Similarity Computing. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 159–163. [\[CrossRef\]](#)

35. Ganitkevitch, J.; Van Durme, B.; Callison-Burch, C. PPDB: The Paraphrase Database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 758–764.
36. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
37. Loper, E.; Bird, S. NLTK: The Natural Language Toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, Barcelona, Spain, 21–26 July 2004; pp. 214–217.
38. Henderson, J.; Merkhofer, E.; Strickhart, L.; Zarrella, G. MITRE at SemEval-2017 Task 1: Simple Semantic Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 185–190. [\[CrossRef\]](#)
39. Shao, Y. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate Semantic Textual Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 130–133. [\[CrossRef\]](#)
40. Al-Natsheh, H.T.; Martinet, L.; Muhlenbach, F.; Zighed, D.A. UdL at SemEval-2017 Task 1: Semantic Textual Similarity Estimation of English Sentence Pairs Using Regression Model over Pairwise Features. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 115–119. [\[CrossRef\]](#)
41. Kohail, S.; Salama, A.R.; Biemann, C. STS-UHH at SemEval-2017 Task 1: Scoring Semantic Textual Similarity Using Supervised and Unsupervised Ensemble. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 175–179. [\[CrossRef\]](#)
42. Lee, I.T.; Goindani, M.; Li, C.; Jin, D.; Johnson, K.M.; Zhang, X.; Pacheco, M.L.; Goldwasser, D. PurdueNLP at SemEval-2017 Task 1: Predicting Semantic Textual Similarity with Paraphrase and Event Embeddings. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 198–202. [\[CrossRef\]](#)
43. Zhuang, W.; Chang, E. Neobility at SemEval-2017 Task 1: An Attention-based Sentence Similarity Model. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 164–169. [\[CrossRef\]](#)
44. Śpiewak, M.; Sobecki, P.; Karaś, D. OPI-JSA at SemEval-2017 Task 1: Application of Ensemble learning for computing semantic textual similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 139–143. [\[CrossRef\]](#)
45. Fialho, P.; Patinho Rodrigues, H.; Coheur, L.; Quaresma, P. L2F/INESC-ID at SemEval-2017 Tasks 1 and 2: Lexical and semantic features in word and textual similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 213–219. [\[CrossRef\]](#)
46. España-Bonet, C.; Barrón-Cedeño, A. Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 144–149. [\[CrossRef\]](#)
47. Bjerva, J.; Östling, R. ResSim at SemEval-2017 Task 1: Multilingual Word Representations for Semantic Textual Similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 154–158. [\[CrossRef\]](#)
48. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning—Volume 32, JMLR.org, ICML’14, Beijing, China, 21–26 June 2014; pp. II-1188–II-1196.
49. Meng, F.; Lu, W.; Zhang, Y.; Cheng, J.; Du, Y.; Han, S. QLUT at SemEval-2017 Task 1: Semantic Textual Similarity Based on Word Embeddings. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 150–153. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.