Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09466-4

ORIGINAL PAPER

A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics

Muhidin Mohamed[1] · **Mourad Oussalah**[2]

Abstract In this paper, we propose a hybrid approach for sentence paraphrase identification. The proposal addresses the problem of evaluating sentence-to-sentence semantic similarity when the sentences contain a set of named-entities. The essence of the proposal is to distinguish the computation of the semantic similarity of named-entity tokens from the rest of the sentence text. More specifically, this is based on the integration of word semantic similarity derived from WordNet taxonomic relations, and named-entity semantic relatedness inferred from Wikipedia entity co-occurrences and underpinned by Normalized Google Distance. In addition, the WordNet similarity measure is enriched with word part-of-speech (PoS) conversion aided with a Categorial Variation database (CatVar), which enhances the lexico-semantics of words. We validated our hybrid approach using two different datasets; Microsoft Research Paraphrase Corpus (MSRPC) and TREC-9 Question Variants. In our empirical evaluation, we showed that our system outperforms baselines and most of the related state-of-the-art systems for paraphrase detection. We also conducted a misidentification analysis to disclose the primary sources of our system errors.

Keywords Paraphrase identification · Named-entity semantic relatedness · WordNet · Wikipedia · Word category subsumption

Much of this work was done while a Ph.D. student at the University of Birmingham.

✉ Muhidin Mohamed
m.mohamed10@aston.ac.uk

[1] Department of Computer Science, EAS, Aston University, Birmingham B4 7ET, UK

[2] Centre for Ubiquitous Computing, Faculty of Information Technology Computer Science, University of Oulu, P.O. Box 4500, 90014 Oulu, Finland

Published online: 16 April 2019

1 Introduction

Paraphrases are sentences conveying the same meaning using alternative language expressions (Dias et al. 2010). The identification of paraphrases is explicitly related to the quantification of the amount of semantic overlap between two textual expressions. This typically involves measuring the extent to which a pair of words, phrases or sentences are semantically related to each other using statistical features from large corpora, e.g., Wikipedia (Taieb et al. 2013) and/or semantic features from knowledge networks such as WordNet (Malik et al. 2007). Paraphrase Identification (PI) is a useful task for many other important NLP applications including Text Summarization, Plagiarism Detection, Intelligent Tutoring Systems, Question Answering, and Machine Translation. For instance, with the use of paraphrase detection, a summarization system can eliminate information redundancy in the extracted summary. Paraphrases can also be used to substantiate the correctness of answers produced by a question answering application. Plagiarism detection is another task that can benefit from PI by identifying texts that have been restated using alternative language. In the case of Intelligent Tutoring systems, one can assess whether students' submissions/answers are semantically equivalent to reference answers exploiting paraphrase identification.

Many of the existing paraphrase detection approaches are substantially built on WordNet taxonomy (Fernando and Stevenson 2008; Kim and Baldwin 2013; Mihalcea et al. 2006; Kozareva and Montoyo 2006; Rus et al. 2008; Das and Smith 2009). WordNet is a lexical database where English words are grouped into sets of synsets and interlinked by means of conceptual-semantic and lexical relations (Miller 1995). Apart from the exploitation of noun and verb hierarchical relations, WordNet enables the construction of useful semantic similarity measures that quantify the extent to which two distinct nouns/verbs are semantically related (Pedersen et al. 2004). This can therefore be extended to phrase and sentence levels allowing the semantic overlap between paraphrases to be established and quantified. Nevertheless, the use of the WordNet-based similarity approach is subject to at least three inherent limitations. Firstly, taxonomic relations are only available for noun and verb classes. Therefore, one can only compute the semantic similarity between a pair of nouns or a pair of verbs. This excludes other PoS categories, such as adverbs and adjectives, from the semantic similarity calculus. Secondly, there is a strong discrepancy between the hierarchies of the noun and verb categories where the noun entity is much more abundant and its associated depth (in the hierarchy) is much more important than that of the verb category (Miller and Hristea 2006). This renders the semantic similarity of the nouns and that of the verb entities somehow biased. Thirdly, many of the common named-entities are absent from the WordNet lexical database (Ponzetto 2010), which, in turn, subsequently reduces the semantic overlap detection capabilities of any WordNet-based similarity measure.

The contributions of this paper are threefold. First, we improved WordNet-based semantic similarity by converting all possible loosely encoded and non-hierarchized word categories (e.g., verbs, adverbs and adjectives) to their corresponding nouns using the CatVar database. This process is referred to as PoS conversion throughout this paper. It allows us to cover a wide range of lexical items that would not have been matched without such conversion. In addition, the choice of nouns as a target word category is motivated by its well-structured full-fledged taxonomy as contrasted with other PoS categories encoded in WordNet. Second, we devised a technique for measuring the semantic relatedness between named-entities by exploring the level of their co-occurrences in Wikipedia articles in the same spirit as Normalized Google Distance (NGD). Third, the PoS conversion enhanced WordNet similarity and the Wikipedia-based named-entity semantic relatedness measures are integrated to form a hybrid system for a comprehensive judgement of paraphrased sentences. The hybrid system combines the collective human knowledge in Wikipedia and the semantic relations between concepts in WordNet to improve sentence paraphrase identification. The proposed approach is next evaluated using a set of publicly available datasets where an extensive comparison with some state-of-the-art approaches has been carried out.

The rest of the paper is structured as follows. Section 2 gives a summary of related works. Section 3 deals with sentence paraphrase detection using WordNet taxonomy highlighting both conventional approach of extending WordNet pairwise semantic similarity to sentence based semantic similarity, and the use of PoS conversion through the aid of CatVar database. Section 4 copes with a new metric introduced for measuring named-entity semantic relatedness using Wikipedia. Section 5 details our hybrid approach for computing the semantic similarity employing both Wikipedia and WordNet. Next, some experimental results are provided in Sect. 6. In Sect. 7, we provide a brief error analysis of the top misclassification sources and draw conclusions in Sect. 8.

2 Related works

Without claiming a full coverage of related literature, we can roughly categorize related works into three high-level categories on the basis of their information source, namely; corpus-based, knowledge-based and hybrid methods.

2.1 Corpus-based methods

First, the application of strategies entirely or substantially based on corpus statistics provided some success in addressing the paraphrase identification problem (Blacoe and Lapata 2012; Ji and Eisenstein 2013; Issa et al. 2018). Ji and Eisenstein (2013) used a simple distributional similarity model by designing a discriminative term-weighting metric called TF-KLD. The authors claimed that their newly introduced metric outperforms the widely used TF-IDF weighing scheme. In the same spirit, Blacoe and Lapata (2012) employed three distributional representations of text: simple semantic space, syntax-aware space and word embeddings. Alternatively, Wan et al. (2006) exploited a machine learning approach using lexical and syntactic dependency-based features whereas other researchers including (Madnani et al. 2012; Finch et al. 2005) investigated the feasibility of WordNet-based machine translation approaches for paraphrase detection. With varying levels of performance on the MSRPC dataset, this category (Corpus-based) contains some of the best-performing methods, notably the works of Ji and Eisenstein (2013) and Issa et al. (2018), which achieved accuracy figures of 80.4% and 86.6%[1] respectively.

2.2 Knowledge-based methods

In the second category, one acknowledges the works of Fernando and Stevenson (2008), who used word level similarities derived from WordNet taxonomy. Similarly, Das and Smith (2009) utilized quasi-synchronous dependency grammars in a probabilistic model incorporating WordNet. Furthermore, Kozareva and Montoyo (2006) advocated an approach based on content overlap (e.g., n-grams and proper names) and semantic features derived from WordNet. Unlike other WordNet-based methods, Hassan (2011) suggested a new approach called Salient Semantic Analysis (SSA) that used context meaning according to Wikipedia links. This class of approaches achieved a relatively subordinate performance of less than 83% in F-measure and below 77% in accuracy on the MSRPC.

2.3 Hybrid methods

The hybrid approaches rely on the use of two or more information sources ranging from distributional statistics, path lengths between concepts in graphical knowledge representations, to more complicated machine learning and feature based algorithms. For instance, Mihalcea et al. (2006) combined corpus-based and knowledge-based semantic similarity using TF-IDF weighted word-to-word maximal similarities derived from WordNet and the British National Corpus. Contrary to the similarity oriented approach, some researchers (e.g., Qiu et al. (2006), Wang et al. (2016)) suggested a paraphrase identification model that considers both the similarity and dissimilarity between sentences. In a more entailment oriented approach, Rus et al. (2008) built a graphical representation of text by mapping relations within its syntactic dependency trees. In another related work, Islam and Inkpen (2008) presented a sentence similarity model based on the semantic and syntactic information. Pairwise semantic features of single words and multiword expressions from syntactic trees have also been utilized in Socher et al. (2011). Recently, there has been a growing interest in applying neural networks to the problem of paraphrase identification (He et al. 2015). The performance of this class of systems is relatively low compared to the previous two groups with the exception of the work of He et al. (2015) who reported accuracy and F-measure results of 78.6% and 84.7%, respectively, on the MSRPC.

2.4 The current work

Our work falls within the realm of hybrid approaches due to its use of combined semantic information issued from Wikipedia corpus, CatVar database and WordNet-derived features. We make use of a semantic similarity approach to

[1] This is in the transductive setting only, their accuracy in the inductive case was 68.7%.

determine the existence of a paraphrase relationship between sentences. Similar to (Fernando and Stevenson 2008; Mihalcea et al. 2006; Kozareva and Montoyo 2006; Rus et al. 2008; Das and Smith 2009), this paper advocates the use of a WordNet-sourced semantics for paraphrase detection. However, several improvements have been put forward in order to address some known WordNet limitations. First, the absence of a hierarchical organization for adjectives and adverbs, and the discrepancy between noun and verb categories have been tackled with the application of PoS transformation using CatVar database (Habash and Dorr 2003). Second, inspired by the NGD rule (Cilibrasi and Vitanyi 2007), the Wikipedia lexical database was employed to derive a new named-entity similarity measure. This is motivated by the continuous expansion of the Wikipedia database and the fact that around 74% of its articles describe named-entities (Nothman et al. 2008).

3 Using WordNet taxonomy for similarity-based paraphrase detection

Prior to word-similarity computation, sentence texts were processed using standard natural language processing packages and parsers including the Illinois PoS and Named-entity Taggers (Ratinov and Roth 2009; Roth and Zelenko 1998) in order to identify the various tokens, their PoS category and the presence of named-entities. The latter is sometimes constituted of composed words (e.g., New York) following the outcome of the named-entity recognizer. Throughout this paper, we confine our reasoning to the commonly employed bag-of-word representation of the aforementioned tokens obtained after applying parsing and named-entity recognition. In this respect, in order to quantify the similarity of two sentences, one distinguishes the conventional WordNet-based approach and alternative approaches developed in this paper.

3.1 Conventional approach

WordNet is a hierarchical lexical database for English developed at Princeton University (Miller 1995). It has four primary word categories: nouns, verbs, adjectives and adverbs. Its words are organized into synsets where each synset contains a number of interchangeable lexical units. Conceptual IS-A relations encoded among synsets create a hierarchical structure from general to more specific concepts, e.g., $researcher^1 @ \Rightarrow scientist^1 @ \Rightarrow person^1 @ \Rightarrow organism^1 @ \Rightarrow livingthing^1$ with $@ \Rightarrow$ and superscripts, respectively, indicating IS-A relations and word senses. This structure provides word sense links which represent semantic information for similarity measures derived from path lengths of knowledge networks. For the WordNet-based word-to-word similarity and relatedness, we used the implementation described in Pedersen et al. (2004). Especially, we considered the common Wu and Palmer (1994) measure, which estimates the semantic relation between two synsets from the position of their concepts, say, c_1 and c_2. It compares the depth of the lowest common subsumer (lcs) of concept 1 and concept 2 ($lcs(c_1, c_2)$) to their total depth from the root node as in Eq. 1.

$$Sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{1}$$

Extrapolating from word semantic similarity measure to sentence similarity requires further investigation as sentences contain a group of words that convey a complete conceptual sense. As such, any means of measuring the semantic similarity between two sentences should use the association from the semantic distance between the concepts where, typically, pairwise comparison of similar word classes using either noun or verb WordNet taxonomy is employed.

With the conventional WordNet approach, the similarity of two words can be computed only if they are of the same PoS and they form part of one of two syntactic categories: nouns or verbs. This is due to the WordNet design in which the adjective and adverb categories lack taxonomic hierarchies. Besides, given that a word may be associated with more than one concept (synset), the semantic similarity between any pair of words is computed from the maximum pairwise conceptual score of the two words. Related studies including (Malik et al. 2007; Mihalcea et al. 2006) applied such a conventional method and extended it to sentence granularity. By this extension, if S_A and S_B denote two sentences to be compared, their semantic similarity, assuming a symmetrical contribution of the two sentences, is computed as per Eq. 2. The word-to-word semantic similarity, *Sim*(w, x), is computed between the same PoS words (PoS(x) = PoS(w)) that are either nouns or verbs.

$$Sim_{WN}(S_A, S_B) = \frac{1}{2}\left(\frac{\sum_{w \in S_A} \max_{x \in S_B} Sim(w, x)}{|S_A|} + \frac{\sum_{w \in S_B} \max_{x \in S_A} Sim(w, x)}{|S_B|}\right) \tag{2}$$

where *Sim*(w, x) stands for the Wu and Palmer WordNet similarity measure, and $|S_A|$ (resp. $|S_B|$) denotes the number of lexical units in sentence A (resp. sentence B). There is an observable anomaly in Eq. 2, especially in the normalization parameter, where the sentence length is used as a normalizing factor. The intuition supports that many words, e.g., named-entities do not appear in WordNet and hence will not contribute to the similarity. In that situation, it makes sense to reflect this in the normalization factor by neglecting all non-contributing words from the sentence length. This will be addressed later on in Eqs. 5, 6, and 7 where each sentence similarity will be normalized by its contributing tokens only.

3.2 An approach aided with word PoS conversion

As shown in Eq. 2, the conventional approach of WordNet semantic similarity is based on averaging over all one-to-one word-level semantic similarities of the two sentences. Nevertheless, this is restricted to pairs of words that belong either to verb or noun PoS categories. Therefore, semantic similarity between words, like *convert* and *conversion* cannot be established in the conventional way because, despite being morphological derivations of the same word, they belong to distinct word

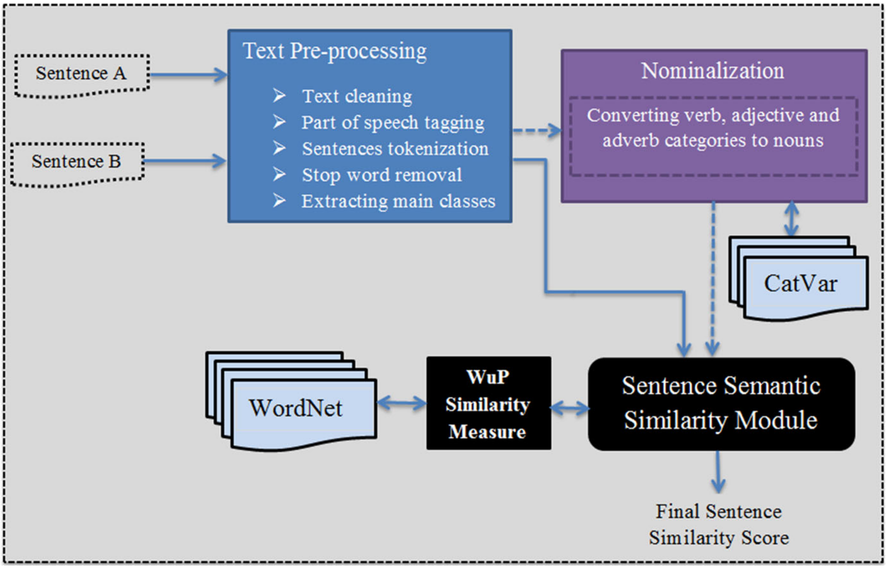

Fig. 1 Sentence semantic similarity assisted with PoS conversion

categories. It also leaves other important sentence tokens, such as proper nouns, adverbs and adjectives unaccounted for (Mohamed and Oussalah 2014).

Subsequently, an approach for addressing the above limitations is developed. It maximizes the sentence semantic space by converting loosely encoded or non-hierarchized word classes into a single strongly hierarchized word category. To this end, three primary word categories, namely verbs, adjectives and adverbs are transformed to their equivalent nouns using CatVar (Habash and Dorr 2003). A block diagram of the proposed CatVar-aided sentence textual similarity measure is depicted in Fig. 1. The system comprises four main modules: Text Pre-processing, Sentence Semantic Similarity, Word PoS Conversion and WordNet Similarity Measure. The Sentence Semantic Similarity module represents the core component of the system. The pre-processed sentence texts are nominalized before being fed into the core sub-system. Note that the sentence similarity can be computed with or without nominalization depending on whether we want to run the proposed PoS conversion aided approach or the conventional method. Example 1 illustrates the system functionality.

Example 1

S_1: “The transformation of word forms is an improvement for the sentence similarity”.
S_2: “Converting word forms enhances the sentence similarity”.

After initial text pre-processing, the two sentences boil down to the following token-based representations with the subscript tags indicating the words’ part-of-speech.

S_1: (transformation$_{NN}$, word$_{NN}$, forms$_{NNS}$, improvement$_{NN}$, sentence$_{NN}$, similarity$_{NN}$).
S_2: (converting$_{VBG}$, word$_{NN}$, forms$_{NNS}$, enhances$_{VBZ}$, sentence$_{NN}$, similarity$_{NN}$).

It is easy to notice that sentence 1, unlike sentence 2, contains no verb PoS, which would result in the verbs *'converting'* and *'enhances'* not contributing to the overall sentence similarity score. However, if a verb-to-noun conversion is applied, *'converting'* will be turned into its equivalent noun *'conversion'*, while *'enhances'* converts to *'enhancement'*. The generated nouns are paired with corresponding nouns from the other sentence, say, *'improvement'* for *'enhancement'* and *'transformation'* for *'conversion'*. Applying Eq. 2 to the nominalised sentences increases the total similarity score from 0.786 to 0.889, which makes it closer to the human intuition as the two sentences are closely related in meaning. Note that maximal word similarities are used when computing the sentence similarity score without paying particular attention to the word order. This is because to find the contribution of each word to the similarity score from the knowledge base, the best matching term (the closest in meaning with the highest score) is selected from the partner sentence irrespective of its position in the text.

3.3 The CatVar-aided PoS conversion algorithm

CatVar is a database of English words containing derivationally-related classes. The categorial variants fall in different parts-of-speech but share the same morphological base form, e.g., *research*$_V$, *researcher*$_N$, and *researchable*$_{AJ}$. Morphological relations are very important for NLP applications. For instance, when determining the semantic similarity between sentences, which typically comprise of different PoS, inter-class transformation can be applied. The lexical database is built in the form of word clusters each containing variations of a particular stem. It was constructed using other lexical resources including WordNet, Longman Dictionary of Contemporary English (LDOCE), the Brown Corpus section of the Penn Treebank, the English morphological analysis lexicon developed for PC-Kimmo (Englex), and NOMLEX (Habash and Dorr 2003).

The CatVar-assisted PoS conversion is accomplished by finding the database cluster containing the word to be nominalized say, *'devote'* and replacing it with the target word *'devotion'* as they are assuredly in the same cluster. We have developed a Perl module that implements the nominalization on this manner using a local Perl readable version of the CatVar database. There were challenges associated with inflectional words, such as nouns in their plural forms or verbs in different tenses during the conversion. Inflectional forms are reduced, after which content morphemes are fed into the PoS converting module. If a CatVar cluster associated with a verb to be nominalized has several equivalent noun alternatives (e.g., *found* which can be converted to any of *founding*, *founder*, *foundation*), the PoS conversion is done as shown in Algorithm 1. The rationale behind this logic is that all nouns in the CatVar cluster are derivationally-related categorial variants of the other open-class words with the gerund considered as the first choice, when applicable. The overall CatVar-aided nominalization process works as follows:

1. For each sentence, we normalize all inflected words with the aid of WordNet lemmatization prior to its CatVar-based nominalization (*e.g, converting ⇒ convert, see Example* 1 *in Sect.* 3.2).
2. Next, all non-noun open-class tokens in the sentence are nominalized to their semantically equivalent noun variants using CatVar database.
3. Finally, we build and return a bag-of-words sentence vector comprising original and converted nouns for each sentence. The output from this algorithm is fed to the WordNet Sentence Similarity Module given in Fig. 1.

Algorithm 1 Nominalization of words with several noun variants

```
1: Input the verb ($v_i$) to be nominalized.
2: Retrieve the cluster associated with $v_i$ from the CatVar database.
3: If the noun forms in the cluster contain a gerund which can be found in
   WordNet, use it as the first noun counterpart of $v_i$.
4: Otherwise pick the first similarity-maximizing noun variant in the cluster as a
   replacement for $v_i$.
```

4 Named entity semantic similarity and relatedness

The word named-entity (NE) as used today in text mining and Natural Language Processing (NLP) was introduced in the Sixth Message Understanding Conference (Grishman and Sundheim 1996). In the context of this work, named-entity refers to the proper names of locations, people, organizations, and other entities (aka miscellaneous). From this definition, a named-entity can be abstract (e.g., Gregorian) or have physical existence (e.g., Barak Obama, Shakespeare). It can also be viewed as entity instances (e.g., New York is an instance of a city, Jaguar is an instance of a car brand). This is typically achieved using named-entity recognition software.

Establishing semantic associations among these named-entities is a critical component in text processing, information retrieval, and knowledge management. Despite this fact and due to the insufficient coverage of these proper names in the language thesaurus and knowledge networks (e.g., dictionaries, WordNet), the accurate determination of the semantic relatedness between two pieces of text containing these entities remains an open challenge and a research problem. For instance, if you search for the world’s largest corporations such as Microsoft and Apple, you are unlikely to find them in the well-established linguistic knowledge resources such as WordNet. Constantly updated online repositories such as Wikipedia, possess a much higher coverage than WordNet in terms of named-entities (Ponzetto 2010; Habib and van Keulen 2016). Therefore, we use Wikipedia utility for named-entity similarity approximation underpinned with the NGD algorithm. Regardless of their type (e.g., location, person, organization, miscellaneous), the semantic relatedness between named-entities is determined using their individual and joint counts in the Wikipedia database.

4.1 Wikipedia entity co-occurrence for named-entity semantic relatedness

In natural languages, some words have a higher probability of co-occurrences than others in language corpora. For example, the name *Joseph S Blatter* is more likely to appear alongside the named-entity *FIFA* than *NASA*. This can be perceived as an indication of the semantic association between the two named-entities. At the time of our experiments, the number of Wikipedia articles containing the names *FIFA* and *Joseph S Blatter* singly were 33,123 and 291 respectively while the Wikipedia pages in which the named-entities occurred jointly were 267, yielding a high similarity score between the two concepts as will be detailed later on. Since its foundation in 2001, Wikipedia has grown in both popularity and size leading to an increased usage among the NLP research community. At the time of our latest experiments, Wikipedia contained over 32 million articles in 260 languages where its English version had more than 5 million articles, containing predominantly well-structured articles. The latter made the encyclopedia to be a reliable resource for any NLP task. Other motivations for the use of the NGD measure on the Wikipedia database for named-entity semantic similarity quantification are summarized below:

1. Empirical and survey research found that around 74% of Wikipedia pages describe named-entities (Nothman et al. 2008) justifying that Wikipedia has a high coverage of named-entities.
2. Current state-of-the-art lexical resources, such as WordNet, provide insufficient coverage of named-entities.
3. Google deprecated its local API access since October 2013 whereas Wikipedia remains publicly open for local access.
4. Our experimental tests, based on NGD via a web interface, showed that Google hits fluctuate over time, suggesting that they are approximate counts.

A given name may sometimes refer to more than one entity triggering the need for an explicit match to be made to the correct instance. That is if several Wikipedia articles contain the same named-entity as their title and a user tries to find it in the database, a potential ambiguity may arise. This is often addressed by the Wikipedia disambiguation pages, which list all possible meanings of the ambiguous entity. However, our current approach does not adopt the Wikipedia disambiguation for two reasons. Firstly, the named-entity component of the proposed hybrid similarity measure relies on the occurrence and co-occurrence counts of the named-entities as their semantic proximity regardless of whether it forms the title or occurs in the article text. That means, when determining the semantic relatedness between two entities, we only need to count the number of Wikipedia articles containing each named-entity, and the figure of articles comprising both entities together. Since the exact names with their actual spelling have to be searched and counted, disambiguation does not seem to be of much help in this case. Secondly, the identities of the names in the original text remain unidentified prior to their retrieval, a process that should have been accomplished before propagating any Wikipedia disambiguation. In any case, adding a disambiguation layer to our current approach can be considered worthwhile, providing room for further improvement.

Our approach for named-entity semantic relatedness is based on entity co-occurrence in the form of Wikipedia article counts underpinned by the NGD rule, a mathematical theory based on Information Distance and Kolmogorov Complexity (Cilibrasi and Vitanyi 2007). Especially, we downscaled NGD to Wikipedia. In other words, if ne_i and ne_j are two entities, we extract the number of Wikipedia articles $AC(ne_i)$, $AC(ne_j)$, & $AC(ne_i, ne_j)$ for the entities ne_i, ne_j and their coexistence respectively. The article counts from Wikipedia are treated as the semantic distance between the two names. More formally, the Wikipedia-based similarity of two named-entities $NWD(ne_i, ne_j)$ can be computed as:

$$NWD(ne_i, ne_j) = \frac{\max\left[log_2 AC(ne_i), log_2 AC(ne_j)\right] - log_2 AC(ne_i, ne_j)}{log_2 N - \min\left[log_2 AC(ne_i), log_2 AC(ne_j)\right]} \tag{3}$$

The parameter N in the denominator is the total number of English Wikipedia articles. Next, inspired by Gracia et al. (2006); Aliguliyev and Aliguliyev (2009), the similarity between named-entities ne_i and ne_j is computed using an exponential function that would guarantee the score to be normalized in the unit interval:

$$Sim_{NWD}(ne_i, ne_j) = e^{-NWD(ne_i, ne_j)} \tag{4}$$

From an implementation perspective, Eq. 4 turns out to be a quite simple, effective and language independent named-entity similarity measure. Additionally, the Wikipedia-based similarity will only be applied if both named-entities possess entries in Wikipedia. This guarantees that $AC(ne_i) > 0$ and $AC(ne_j) > 0$, and thereby, expressions (3–4) are always fully defined. The approach can also be employed for common open-class words, not necessarily named-entities, provided the existence of a Wikipedia entry. But such an approach has not been pursued in this paper, although one acknowledges other related works following such direction (Gabrilovich and Markovitch 1996; Taieb et al. 2013; Mohamed and Oussalah 2016). The application of Eqs. 3 and 4 to seek the semantic similarity between named-entities FIFA and Sepp Blatter, with previously indicated article counts, yields:

$$\begin{aligned} Sim_{NWD}(FIFA, Sepp\,Blatter) &= e^{-NWD(FIFA, Sepp\,Blatter)} \\ &= e^{-\frac{\max\left[log_2(33123), log_2(291)\right] - log_2(267)}{log_2 4617085 - \min\left[log_2(33123), log_2(291)\right]}} \\ &= e^{-0.4984} \approx 0.6075 \end{aligned}$$

Equation 4 can also be extended to determine the sentence-to-sentence semantic similarity in view of their named-entities only. Namely, let us say that NE_1 represents the set of named-entities contained in the first sentence and NE_2 the set of named-entities in the second sentence, then the associated semantic relatedness is calculated as:

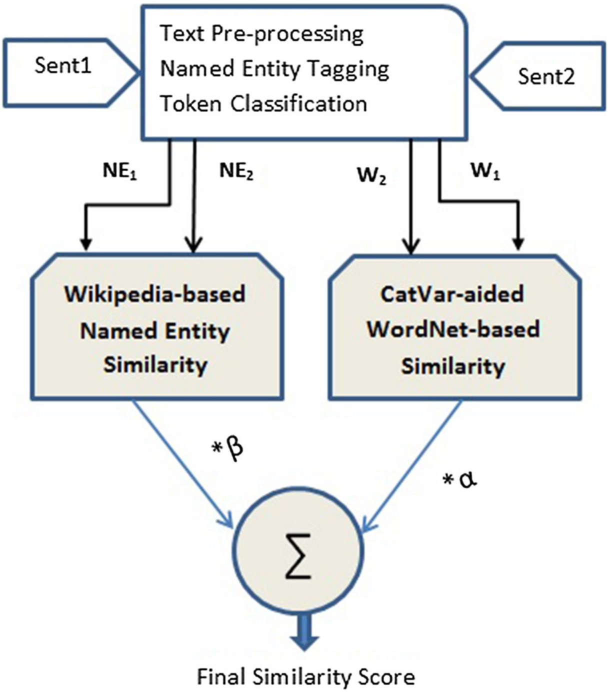

Fig. 2 Hybrid system for sentence paraphrase detection

$$Sim_{WP}(NE_1, NE_2) = \frac{1}{2}\left(\frac{\sum\limits_{ne_i \in NE_1} \max\limits_{ne_j \in NE_2} Sim_{NWD}(ne_i, ne_j)}{|NE_1|} + \frac{\sum\limits_{ne_j \in NE_2} \max\limits_{ne_i \in NE_1} Sim_{NWD}(ne_i, ne_j)}{|NE_2|}\right) \tag{5}$$

Equation 5 assumes a similar contribution of both sentences to the similarity score in the same spirit as Mihalcea et al. (2006). Especially, if the two sentences contain a single named-entity each, then (5) coincides with (4). Trivially, if the two sentences have named-entities which have high similarity scores in the sense of $Sim_{NWD}(ne_i, ne_j)$ for each ne_i of the first sentence and ne_j of the second sentence, then straightforwardly, the resulting $Sim_{WP}(NE_1, NE_2)$ is equally high. “Appendix 1” summarizes some interesting properties of the proposed named-entity semantic relatedness measure.

5 The proposed hybrid method

Figure 2 shows the hybrid system. It is an integration of the CatVar-enhanced WordNet similarity and Wikipedia-based named-entity similarity through some convex combination of the two inputs. We achieved the system implementation with Perl scripts in a Linux environment. For the Wikipedia based similarity component, we extracted Wikipedia article counts associated with named-entities by parsing the raw Wikipedia entries retrieved via a custom search. Specifically, we performed the search for the entities and counted their occurrences in the Wikipedia knowledge base through a web interface. The mechanism of the interface is built on Wikipedia Automated Interface[2] (Summers and Cassidy 2011), which enables the system to search and extract Wikipedia pages. Once recovered, the articles are parsed and pattern-searched using regular expressions to allow the enumeration of articles containing the named-entities being considered. The joint counts, which are used in Eq. 3, imply semantic proximity between the named-entities. As for the word level similarity of the WordNet-based component, we adapted the implementation of WordNet similarity measures (Pedersen et al. 2004) for computing conceptual relatedness of individual words after applying the CatVar-aided PoS conversion.

In addition to the typical text pre-processing steps (e.g., sentence splitting, tokenization, stop-word removal), two more system specific tasks; namely, named-entity tagging and token classification have been applied to the input texts. Named-entity tagging is the process of recognizing and labelling all proper nouns in the text (Grishman and Sundheim 1996). Also, token classification is a post tagging step in which sentence tokens are split into content word and named-entity vectors. In Fig. 2, the inputs to the subsystems denoted by the notations NE_1, NE_2, W_1, and W_2 are all term vectors of the corresponding sentence with NE_1 and W_1 being the named-entity and common word vectors for sentence 1. A generic formula for the semantic similarity of non-named-entity sets W_1 and W_2 yields

$$Sim_{WN}(W_1, W_2) = \frac{1}{2}\left(\frac{\sum_{w_i \in W_1} \max_{w_j \in W_2} Sim(w_i, w_j)}{|W_1|} + \frac{\sum_{w_j \in W_2} \max_{w_i \in W_1} Sim(w_i, w_j)}{|W_2|}\right) \quad (6)$$

Finally, the overall semantic similarity of the two sentences, taking into account the occurrence of named-entities and non-named-entities is given as the convex combination of the Sim_{WN} and Sim_{WP}:

$$Sim(S_1, S_2) = \alpha Sim_{WN}(W_1, W_2) + \beta Sim_{WP}(NE_1, NE_2) \quad (7)$$

The coefficients α and β ($0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $\alpha + \beta = 1$) balance the contribution of the Wikipedia-based and WordNet-based similarity components.

A simple modelling of the convex coefficients relies on the number of entity and word tokens employed in Wikipedia-based and WordNet-based similarity components. This follows the statistical argumentation that the more the number of tokens associated to WordNet is higher than the number of named-entities in both sentences, the more one expects the contribution of Sim_{WN} to be more important than that of Sim_{WP} in the hybrid model. More specifically, let $W_1,\ W_2$ be the set of WordNet related tokens in the first and second sentence, respectively. Let NE_1 and

[2] https://metacpan.org/release/WWW-Wikipedia.

NE_2 be the set of named-entities in the first and second sentence, respectively. Then the parameters α and β can be given as:

$$\alpha = \frac{|W_1| + |W_2|}{|W_1| + |W_2| + |NE_1| + |NE_2|}; \quad \beta = \frac{|NE_1| + |NE_2|}{|W_1| + |W_2| + |NE_1| + |NE_2|} \tag{8}$$

The use of word proportions from the sentence pairs (Eq. 8) as coefficients for the combination of the two similarity components (Eq. 7) has some desirable attributes. First, it conforms with unity sum, and second, it serves as a weighting control strategy for the relative contribution of each similarity component. For instance, in the boundary case of Eq. 8, it is easy to see that if there are no named-entities in the pair of sentences, then $|NE_1| = |NE_2| = 0$, which entails $\alpha = 1 \,\&\, \beta = 0$, so that $Sim(S_1, S_2) = Sim_{WN}(W_1, W_2)$. Similarly, if the pair of sentences are primarily constituted of entities, then $\beta = 1 \,\&\, \alpha = 0$ which entails $Sim(S_1, S_2) = Sim_{WP}(NE_1, NE_2)$. Even in the case where only one sentence contains a named-entity (resp. non-named-entity token), the system ignores that occurrence as the Wikipedia-based similarity can only be performed if named-entities in both sentences possess entries in the Wikipedia database (resp. existence of noun counterpart in the other sentence).

5.1 Illustrative examples

Examples 2, 3 and 4 illustrate the functioning of the hybrid approach, and show its advantages over either individual WordNet-based or Wikipedia-based similarity. For all examples, the similarity of each word (noun or a named-entity) is taken from the best matching word (highest similarity score interpreted as the closest in meaning) of the other sentence following one-to-one pairwise comparison of all terms in the text.

Example 2 (Color text online)

Sent1: Joseph Chamberlain was the first chancellor of the University of Birmingham.
Sent2: Joseph Chamberlain founded the University of Birmingham.

The limitations pointed out for WordNet only based semantic similarity are clearly observable in this example as neither *'chancellor'* nor *'founded'* can be quantified due to the absence of similar PoS word in the partner sentence. Similarly, the two compound named-entities, *'Joseph Chamberlain'* and *'University of Birmingham'* in both sentences, are not covered in WordNet. For simplicity, both sentences have three tokens each: two named-entities and one common word. If we assume that Sent1 tokenizes to $S_1 = (ne_{11}, ne_{12}, w_{11})$ and Sent2 tokenizes to $S_2 = (ne_{21}, ne_{22}, w_{21})$, then we can place each sentence across the columns or the rows as in Table 1. In this representation, the order of the sentences is preserved, i.e., both ne_{11} and ne_{21} stand for *'Joseph Chamberlain'*, ne_{12} and ne_{22} denote *'University of Birmingham'*, while w_{11} and w_{21} represent the words *'chancellor'* and *'founded'*. Table 1 presents the pairwise word comparisons for conventional WordNet, WordNet with CatVar conversion and the proposed hybrid method.

Table 1 Pairwise token comparison of Example 2 using different similarity measures

Sent$_2$	*Sent*$_1$			
	ne_{11}	w_{11}	ne_{12}	Max
(A) Conventional WordNet similarity				
ne_{21}	0^*	0^*	0^*	0
w^*_{21}	0^*	0^*	0^*	0
ne_{22}	0^*	0^*	0^*	0
Max	0	0	0	**0***
(B) CatVar-aided WordNet similarity				
ne_{21}	0	0	0	0
w^*_{21}	0	0.19	0	0.19
ne_{22}	0	0	0	0
Max	0	0.19	0	**0.19***
(C) Hybrid scheme similarity				
ne_{21}	1	0	0.49	1
w^*_{21}	0	0.19	0	0.19
ne_{22}	0.49	0	1	1
Max	1	0.19	1	**0.76***

From Table 1(A), all word pairings of the conventional WordNet similarity yield zero scores (0*) as the included named-entities are not covered in WordNet and the only two common words differ in PoS. In Table 1(B), a nominalization is incorporated which means that all verbs (*founded* only in this case) are turned to nouns. In addition to applying word PoS conversion, Wikipedia-based named-entity similarity is augmented to form the hybrid method as given in Table 1(C). Maximum scores of each row and column are listed in the corresponding cells. The highlighted value in the last cell of every row and column in each of the three subtables is the final similarity score of the respective scheme as per Eqs. (2, 5, 6, 7). Improvements achieved through the single word PoS conversion ($0 \rightarrow 0.19$) and further page count retrieval of the two proper nouns from Wikipedia ($0.19 \rightarrow 0.76$) are already apparent through the obtained scores.

Noticeably, the two shared named-entities in Example 2 are string identical and exist in both sentences, which is predominantly the case for paraphrases. That is, if a sentence contains a named-entity, all other paraphrases constructed from that particular sentence are highly likely to contain the same entity. Nevertheless, it should be noted that the proposed Wikipedia-based measure also works for paraphrases where each sentence bears different named-entities, or an entity with two or more referent names. Example 3 is a good illustration of this situation where the Brazilian football legend *Edson Arantes do Nascimento* is referred to by his popular nickname (Pele) in the first sentence and by his full name in the second. Additionally, the other two named-entities in the example differ in their letter strings, albeit their strong semantic relatedness. Table 2 shows word similarity scores of the sentences using the same procedure and terminology as for Table 1. Example 3 demonstrates how the Wikipedia-based approach measures the

Table 2 Pairwise token comparison of Example 3 using CatVar-aided WordNet and hybrid similarity measures

$Sent_2$	$Sent_1$					
	ne_{11}	w_{11}	w_{12}	w_{13}	ne_{12}	Max
(A) CatVar-aided WordNet similarity						
ne_{21}	0	0	0	0	0	0
w_{21}	0	0.32	0.7	0.7	0	0.7
w_{22}	0	0.67	1.0	0.69	0	1
w_{23}	0	0.21	0.7	0.7	0	0.7
ne_{22}	0	0	0	0	0	0
Max	0	0.67	1.0	0.7	0	**0.48***
(B) Hybrid scheme similarity						
ne_{21}	0.67	0	0	0	0.63	0.67
w_{21}	0	0.32	0.7	0.7	0	0.7
w_{22}	0	0.67	1.0	0.69	0	1
w_{23}	0	0.21	0.7	0.7	0	0.7
ne_{22}	0.71	0	0	0	0.67	0.71
Max	0.71	0.67	1.0	0.7	0.67	**0.75***

similarity of named-entities that are either not string identical and/or refer to the same real-world object. Different from the previous example, each sentence of the current example has three content words (in blue) denoted in Table 2 by (w_{i1}, w_{i2}, w_{i3}) and two named-entities (in red) symbolized as (ne_{i1}, ne_{i2}), where $i\,(i = 1, 2)$ represents the sentence number. The results in the table reaffirm the ability of the Wikipedia-based metric to measure the semantic relatedness between names in the encyclopedia, be they distinct or identical in their spelling.

Example 3 (Colour text online)

Sent1: Pele penned his first football contract with Santos FC.
Sent2: Edson Arantes do Nascimento started his football career in Vila Belmiro.

The case where the same entity has two or more different names, e.g., *'Pele'*, warrants some attention. Such distinct names for the same referent pose a challenge in which one reference may be more popular, hence more frequent than the other in the Wikipedia corpus. For example, there are 33, 3585 and 28 Wikipedia articles containing *Edson Arantes do Nascimento*, *Pele*, and their combination respectively, indicating an enormous difference in their distribution in the corpus. The preceding implies that the less popular name (*'Edson Arantes do Nascimento'* is rarely used to refer to the entity, which negatively impacts the similarity score because of the imbalance of the page counts. Clearly, factors such as the popularity of the names and their likely use by the Wikipedia contributors, when composing or editing articles, affect the similarity calculation, as seen in this example. One possible way to circumvent this could be through following the Wikipedia redirections, whenever available, which can serve as a means of entity linking and may result in a single reference for the entity. However, this can only be applied if there is a prior

Table 3 Pairwise token comparison of Example 4 using the hybrid similarity measure

Sent$_2$	*Sent*$_1$			
	ne_{11}	w_{11}	ne_{12}	Max
Based on the hybrid similarity measure				
ne_{21}	0	0	0*	0
w_{21}	0	0.57	0	0.57
ne_{22}	0.22	0	0	0.22
Max	0.22	0.57	0	**0.27***

determination of the exact referent of the named-entity (disambiguation), a topic that remains unaddressed in the current work.

Example 4 (Colour text online)

Sent1: Tower Bridge was designed by Sir Horace Jones.
Sent2: Charles Ranlett Flint established IBM.

Another situation worth illustrating is where two negative paraphrases contain non-identical named-entities which are also semantically unrelated. This case is illustrated in Example 4 in which the only two common nouns to be paired after nominalization, unlike the named-entities, are semantically related. Applying the proposed hybrid scheme, while using the same procedure and steps for computing the sentence similarity in Example 4 as for the previous two examples, yields an overall normalized similarity score of *0.27*. The final score is predominantly the contribution of the two words *'designed'* and *'established'* (cell w_{11}–w_{21} in Table 3). This shows that the proposed measure accurately captures the impact of the unrelated named-entity similarity component which hardly contributes to the similarity score on this occasion.

6 Experiments

6.1 Datasets

In our system experiments, we used two different datasets, namely Microsoft Research Paraphrase Corpus and TREC-9 Question Variants, both of which are briefly described below.

6.1.1 Microsoft Research Paraphrase Corpus

Microsoft Research Paraphrase Corpus (MSRPC) is a human annotated dataset created from news articles on the web for the evaluation of machine-based paraphrase identification tasks (Dolan et al. 2004). Its creation has undergone a series of refining stages from which developers finally produced a set of 5801 sentence pairs. The data is unequally split into 30% testing and 70% training. We used 750 sentence pairs (32% negatives, 68% positives) extracted from the training

Table 4 Notations for different similarity measures

CosSim	Cosine similarity
WNwoC	WordNet without conversion
WNwCC	WordNet with CatVar conversion
NeSim	Wikipedia-based entity similarity
Hm	Hybrid method

Table 5 System-baseline comparison on the TREC-9 dataset

Measure	Precision	Recall	F-measure	Accuracy
WNwoC	0.974	0.639	0.772	0.676
CosSim	0.979	0.395	0.563	0.475
WNwCC	0.978	0.731	0.837	0.755
NeSim	**1**	0.647	0.786	0.698
Hm	0.808	**1**	**0.894**	**0.871**

data to determine an optimum demarcation threshold for the classification. For the performance evaluation, we used the entire test data (1725 pairs) consisting of 33.5% false and 66.5% true paraphrases.

6.1.2 TREC-9 Question Variants

Similar to the MSRPC, the TREC-9 Question Variants dataset[3] is created by human assessors to describe semantically identical but syntactically different questions. The dataset contains 54 sets, each derived from an original question paraphrased to equivalent variants ranging from 1 to 7 questions. Unlike the MSRPC, it is characterised by a smaller size and shorter sentence lengths. We created 228 sentence pairs from the same dataset classified into two groups: semantically equivalent composed of an original question and its paraphrased variants, and dissimilar questions randomly paired from its different subsets. The proportion of positive paraphrases in this dataset were slightly higher compared to the MSRPC with 78% and 85%, respectively, representing the real paraphrases for training and testing, while the remaining parts constituted the negative paraphrases.

6.2 Results and discussion

The significance of named-entities in the used datasets is shown by the fact that more than 71% of the paraphrase pairs contain one or more named-entities in both the TREC-9 and MSRPC datasets. This is more supportive evidence which signifies the importance of these textual components, often underestimated in the state-of-the-art knowledge-based similarity approaches. Although, one appreciates other recent related methods where named-entities are not particularly emphasized but treated as any other semantic words in the text (Kusner et al. 2015; Mohamed and

[3] http://trec.nist.gov/data/qa/t9_qadata.html.

Table 6 System-baseline comparison on the MSRPC dataset

Measure	Precision	Recall	F-measure	Accuracy
WNwoC	0.826	0.559	0.667	0.558
CosSim	**0.907**	0.314	0.466	0.432
WNwCC	0.818	0.802	0.810	0.703
NeSim	0.794	0.559	0.656	0.537
Hm	0.820	**0.887**	**0.852**	**0.757**

Oussalah 2016). Empirically speaking, the higher the number of named-entity tokens in a sentence pair (i.e., the more the Wikipedia-based named-entity semantic similarity is weighted), the better the performance of the paraphrase detection in terms of its recall, accuracy and F-measure. This might be due to the nature of named-entities that preserve their spelling regardless of the paraphrasing while content words are either changed or replaced by new ones. For instance, in the pair (What kind of animal was Winnie the Pooh?/What was the species of Winnie the Pooh?), the name *Winnie the Pooh* has the same form in both questions while the common word *kind* gets paraphrased to *species*.

The primary focus of our experiments is on the evaluation of the hybrid method, which determines if two given sentences are negative or positive paraphrases. However, prior to the combined method (*Hm*), we performed a reinforcing assessment of the conversion aided WordNet semantic similarity (*WNwCC*) and the Wikipedia-based named-entity semantic relatedness (*NeSim*) schemes separately. This is to give an indication of the performance of each sub-system in isolation and the substantial improvement achieved after their combination. Evaluation results of these systems along with baselines are given in Tables 5, 6 for the TREC-9 and MSRPC corpora respectively, while related notations are defined in Table 4.

Initially, we ran a set of training experiments using 750 sentence pairs from MSRPC and 30% of the total TREC-9 dataset. During this training, we determined a value of 0.7 to be the threshold that jointly optimises both F-measure and accuracy. In other words, we classify sentence pairs as true paraphrases if their overall semantic similarity score equals or exceeds *0.7*. All other pairs whose similarity scores are less than the threshold are identified as negative paraphrases. One attractive property of using a high threshold is that it reduces the probability of misidentifying negative paraphrases with significant semantic overlaps whereas a low threshold can easily and mistakenly identify these negative paraphrases as semantic equivalents.

Next, we selected two similarity measures; namely, cosine similarity (*CosSim*) and conventional WordNet (*WNwoC*) as baselines. Cosine similarity quantifies the similarity between two pieces of text in the form of word vectors (aka bag of words—BoW). The *CosSim* measure is implemented using BoW model and TF-IDF weighting while conventional WordNet is as explained in Sect. 3.1. These two benchmark methods are evaluated against our proposed conversion-aided WordNet, the Wikipedia-based and the hybrid methods (Tables 5, 6). Notably, the system's better performance on the TREC-9 dataset, as in Table 5, might be due to either the dominance of named-entities after the elimination of stop words, and/or its smaller

Table 7 Comparing our results to relevant state-of-the-art unsupervised PI methods on the MSRPC dataset

System	F-measure (%)	Accuracy (%)
Mihalcea et al. (2006)	81.3 (4)	70.3 (6)
Islam and Inkpen (2008)	81.3 (4)	72.6 (3)
Fernando and Stevenson (2008)	82.4 (2)	74.1 (2)
Rus et al. (2008)	80.5 (5)	70.6 (5)
Hassan (2011)	81.4 (3)	72.5 (4)
Our work	**85.2 (1)**	**75.7 (1)**

Table 8 Comparing our results to relevant state-of-the-art supervised PI methods on the MSRPC dataset

System	F-measure (%)	Accuracy (%)
Finch et al. (2005)	82.7 (7)	75.0 (9)
Wan et al. (2006)	83.0 (6)	75.6 (8)
Das and Smith (2009)	82.7 (7)	76.1 (6)
Socher et al. (2011)	83.6 (5)	76.8 (5)
Blacoe and Lapata (2012)	82.3 (8)	73.0 (9)
Madnani et al. (2012)	84.1 (4)	77.4 (4)
Ji and Eisenstein (2013)	**85.96 (1)**	**80.41 (1)**
He et al. (2015)	84.7 (3)	78.6 (2)
Wang et al. (2016)	84.7 (3)	78.4 (3)
Our work	85.2 (2)	75.7 (7)

size and shorter sentence lengths as compared to the MSRPC corpus. Interestingly, the Wikipedia-based named-entity similarity measure can reliably achieve near WordNet performance, which, in turn, indicates the significance of designated names in a full-text semantic extraction. Therefore, it is not surprising for the combined approach to show better performance in comparison to the separate sub-systems.

We also used McNemar's test to determine whether the improvements attained with the proposed methods are statistically significant in comparison to the baselines. McNemar test tells us whether two classification methods have the same error rate at a given significant level, e.g., $\alpha = 0.05$. The test results showed that the hybrid similarity measure (Hm) significantly improved the paraphrase identification performance compared to *CosSim* ($p < 0.001$), *WNwoC* ($p < 0.001$) and *WNwCC* ($p < 0.005$). The CatVar-aided WordNet (*WNwCC*) also achieved significantly better performance compared to *CosSim* and *WNwoC* (both $ps < 0.001$).

6.2.1 Comparison with related works

As presented in Tables 5, 6, the system-baseline comparison indicated that the CatVar-aided and the hybrid methods outperformed the baselines. Furthermore, we performed an additional evaluation step by comparing our system's paraphrase detection level with related state-of-the-art works for paraphrase identification. To this end, we compare our results with two categories of PI systems, namely unsupervised (Table 7) and supervised (Table 8) approaches. All paraphrase identification methods used to compare our system are based on the MSRPC dataset. Consequently, only the MSRPC results can be considered for strict comparison, which is why we excluded the TREC-9 results from Tables (7, 8). Notably, our unsupervised knowledge-enriched heuristic method stands out among its class of systems in Table 7. It also ranks second in the F-score when compared with the supervised approaches as shown in Table 8. The numbers in the parenthesis following the F-measure and the accuracy values in the tables indicate the ranking of each system in the list.

Of the state-of-the-art comparators in Tables 7, 8, the works of Madnani et al. (2012), Finch et al. (2005), Mihalcea et al. (2006), and Fernando and Stevenson (2008) are more closely related to the current work in terms of the implementation method, the semantic features and the applied external resource. Two of the systems; namely, Fernando and Stevenson (2008) and Mihalcea et al. (2006), are the most relevant of the four to our work as they substantially make use of word similarities and corpus information. The third and fourth systems, Madnani et al. (2012); Finch et al. (2005), are based on machine translation metrics in which word-sequence overlaps and lexical matches are used for similarity scoring. All the four studies used WordNet semantic information, but their performances are worse than the proposed approach excluding the work of Madnani et al. (2012) which achieves better accuracy. The improvement of the current approach over these systems is thought to be linked to both CatVar-aided subsumption of non-noun open-class words under derivationally related nouns in WordNet taxonomy and its enrichment with Wikipedia-based named-entity semantic relatedness.

Overall, from Tables 5, 6, 7, 8, it is evident that the combination of Wikipedia and WordNet has clearly improved the paraphrase identification performance. That is to say, using an algorithmically simpler approach, our results outperform all related unsupervised works and improve performance aspects (e.g., F-scores) of the present state-of-the-art supervised systems. This clearly advocates the utilization of WordNet noun taxonomy and its enrichment with named-entity rich resources, such as Wikipedia, for sentence textual similarity and paraphrase identification applications.

7 Misidentification analysis

From what we have presented in Sect. 6.2, the proposed system fails to detect a fair portion of sentence pairs from both datasets with the observation of slightly noticeable better performance on the TREC-9 Question Variants. This may be due to the significant difference in size between the TREC-9 and the MSRPC datasets. In general, we have observed that a significant portion of the system misclassifications come from the high number of generated false positives for different reasons. The following sections briefly discuss the main scrutinized sources of the encountered errors without the exclusion of other possible causes for both TREC-9 and MSRPC datasets. It is noteworthy that although the causes of errors listed under each dataset are primarily linked to that dataset according to our analysis, one should not assume that they cannot apply to the other dataset at all.

7.1 TREC-9 Question Variants

7.1.1 One of the sentence components does not contribute

TREC-9 questions are usually very short in length, where some of the created pairs will only contain named-entities following the preprocessing of their texts. For example the paraphrased questions: *Who was Jane Goodall? and Why is Jane Goodall famous?* will reduce to *Jane Goodall* and *Jane Goodall famous* for the first and the second questions in order. As the first question does not contain any content words, the word *'famous'* in the second will not contribute to the similarity. On the other hand, the absence of named-entities from one or both pairs will also lead to the same error as the contribution of the respective similarity will yield zero.

7.1.2 Part-of-speech tagging errors

The dataset has undergone preprocessing tasks including PoS tagging. Although different part-of-speech taggers can achieve different levels of tagging accuracy with various datasets, they generally introduce errors by incorrectly tagging some of the sentence tokens which consequently lead to system misidentifications. For instance, based on the used Illinois Part-of-Speech Tagger (Roth and Zelenko 1998), the question pair: What date is Boxing Day?; Boxing Day is celebrated at what date? is tagged as *What/WP date/NN is/VBZ Boxing/VBG Day/NNP.?; Boxing/NNP Day/NNP is/VBZ celebrated/JJ on/IN what/WP date/NN.?*. The words *Boxing* and *celebrated* are incorrectly given the wrong tags. Accordingly, this hinders subsequent linguistic manipulations and undermines the correct paraphrase identification.

7.2 Microsoft Research Paraphrase Corpus

7.2.1 Named-entity overlap in non-paraphrase pairs

It appears that the MSRPC's human annotators were not strictly consistent in specifying the level of semantic overlap at which a pair of sentences can be declared unrelated. We have scrutinized that a large proportion of negative paraphrases contain a significant overlap particularly in terms of shared named-entities. For instance, the pair, **Ballmer** has been vocal in the past warning that Linux is a threat to **Microsoft.**; In the memo, **Ballmer** reiterated the open source threat to **Microsoft.,** has been rated as semantically unrelated with such a high observable relevance, particularly of the shared names. Our Wikipedia-based named-entity relatedness measure boosts the pair's similarity score by capturing such a high named-entity overlap. As a result, the hybrid system misidentifies such negative paraphrases as positive paraphrases.

7.2.2 Non-mutual entailments

Actual paraphrases exhibit bidirectional entailments, however, some negative MSRPC paraphrase pairs render a unidirectional entailment where one sentence includes the other with additional information, e.g., *There are 103 Democrats in the Assembly and 47 Republicans; Democrats dominate the Assembly while Republicans control the Senate*. Such pairs possess considerable lexical and semantic overlap and is misclassified as paraphrases by the system while the human raters judged them as non-semantic equivalents.

7.2.3 Ambiguity and coverage of named-entities

A further examination of the system misidentifications showed that the ambiguity of some named-entities in Wikipedia hindered the accurate determination of their semantic relatedness. In other words, the named-entities are not available in the Wikipedia database by their present surface form as in the MSRPC dataset due either to their ambiguity or to the lack of coverage in Wikipedia. One example is the positive paraphrase pair: *Pappas said he wouldn't hesitate about asking Graham to substitute.; Pappas, the teacher, said he wouldn't hesitate having Graham as a substitute*. **Pappas** and **Graham** are ambiguous shortened names which may refer to many people. Disambiguating such names and linking them to their full names may have solved this problem; however, this is one of the study's limitations which might be considered in the future.

7.2.4 Named-entity tagging errors

We have adopted the Illinois Named-entity Tagger (Ratinov and Roth 2009) to recognise and label the four classic types of named-entities: *people, organizations, locations,* and *miscellaneous*. Like other state-of-the-art taggers which use corpus extracted gazetteers, it fails to properly tag some named-entities or labels them with the wrong tags. Mislabelling named-entities and open-class words, as described in Sect. 7.1.2, has been a contributing factor of the system's paraphrase misidentifications.

7.3 Other errors related to the limitations of the study

In addition to the causes of errors described in the previous sections, paraphrase misclassification can also originate from other system limitations. For example, the hybrid approach may fail to correctly classify in the case of false paraphrase pairs where negation is applied to construct the paraphrase, i.e., when using negative particles such as *not* or its reduced form (**n't*), because these terms always form part of the standard stop words. To illustrate that, take the simple pair: *I want a hot breakfast.; I don't want a hot breakfast*. By pre-processing the two sentences and dropping the negative stop word (don't), both sentences reduce to the three terms; *want, hot, breakfast*, which results in a similarity score of 1 when compared, causing the pair to be misjudged as a true paraphrase.

Another situation in which the performance of the system can be undermined is when two paraphrased sentences contain a reference to the same entity, but is referred to using two distinct expressions, e.g., a common noun/pronoun in one sentence and a named-entity in the other sentence. For instance, the final similarity score of the paraphrase pair: *Angela Merkel wants to stand for a fourth election; the chancellor seeks fourth term in office*, can be improved if *Angela Merkel* and *chancellor* are determined to be referring to the same entity. Then, only one of the terms will be used and self-paired inducing the final similarity score to be raised. Identifying such mentions of the same entity in the text (aka coreference resolution) may help overcome this weakness, providing another avenue for future work.

8 Summary and conclusion

We described a hybrid sentence paraphrase identification approach. The primary goal of this approach is to study how the combination of WordNet-based similarity, enriched with CatVar-aided nominalization, and crowdsourced encyclopaedic knowledge in Wikipedia augments the performance of paraphrase identification. To this end, we maximized the comparable semantic tokens by subsuming three primary word categories of verbs, adverbs, and adjectives under derivationally related nouns in WordNet taxonomy. The word class subsumption (PoS conversion) is performed using CatVar database. Changing the part-of-speech of words achieved tangible improvement of sentence paraphrase detection. The performance is further improved with the use of Wikipedia as an external knowledge repository for named-entities. In the combined approach, each sentence is partitioned into two semantic vectors, content words and named-entities. The similarity of the content word vectors is computed from WordNet taxonomy whereas the semantic relatedness of named-entities is based on Wikipedia article counts underpinned with NGD. Some properties of the hybrid method have been investigated (cf. "Appendix 1"). The proposal has been applied to the two publicly available datasets of Microsoft Research Paraphrase Corpus and the TREC-9 Question Variants. The obtained experimental results show that our system outperforms baselines and most of the state-of-the-art systems for sentence paraphrase detection.

There are some system limitations in which addressing them can form potential avenues for future work and may improve the proposed approach. These include, among others: (1) adding a disambiguation step to link all ambiguous named-entities to their actual referents prior to computing their semantic relatedness; (2) employing a strategy for coreference and anaphora resolution to optimize word pairing and similarity computation, particularly in situations where all references of the same entity can be replaced with a single name; (3) exploring a technique for penalizing or down-weighting the similarity scores of negative paraphrases where negation is used to construct the paraphrase to avoid paraphrase misclassification; (4) examining textual similarity using semantic role labeling to address the issues of word contexts within sentences, and consideration of word syntactic order and semantic roles.

Appendix 1

Discussion on the proposed named entity similarity measure (see Sect. 4)

Equations (3–5) deserve special attention when looking at their boundary condition and monotonicity behaviour:

- Assuming the similarity function (4) as inducing a relation between two named-entities, say, $ne_i \Re ne_j$ if and only if $Sim_{NWD}(ne_i, ne_j) \geq \delta$ (δ is some threshold value, $0 < \delta \leq 1$), then it is easy to see that $\Re$ is ***reflexive***, e.g., for any identical named-entities, it holds $Sim_{NWD}(ne_i, ne_i) = 1$, ***symmetric*** because of the symmetry of Sim_{NWD} (e.g., $Sim_{NWD}(ne_i, ne_j) = Sim_{NWD}(ne_j, ne_i)$). However, $\Re$ is not ***transitive***, as it is easy to find three named-entities in Wikipedia such that $Sim_{NWD}(ne_i, ne_j) \geq \delta$ and $Sim_{NWD}(ne_j, ne_l) \geq \delta$ but $Sim_{NWD}(ne_i, ne_l) < \delta$. Nevertheless, it should be noted that if a weaker construction of $\Re$ is allowed, where more flexibility in terms of the definition of the threshold δ is enabled, then the transitivity can be restored. This follows from the observation that if there is co-occurrence of named-entities ne_i and ne_j, and between ne_j and ne_l, then predominantly, there is also co-occurrence between named-entities ne_i and ne_l, although, not necessarily on the same order of magnitude to ensure the strict fulfillment of the transitivity relation (for sufficiently high value of δ).
- If there are no ***co-occurrences*** of named-entities ne_i and ne_j in Wikipedia, then $AC(ne_i, ne_j) = 0$. Substituting this into Equation 3 yields $NWD(ne_i, ne_j) = +\infty$. Therefore, $Sim_{NWD}(ne_i, ne_j) = 0$. Besides, it is easy to see from (3) that $NWD(ne_i, ne_j) = +\infty$ entails $AC(ne_i, ne_j) = 0$. This indicates that the Wikipedia-based similarity is minimal for any pair of named-entities who do not co-occur. In contrast, if the occurrence of named-entity ne_i always coincides with an occurrence of named-entity ne_j, e.g., any Wikipedia article containing ne_i also contains ne_j, then $AC(ne_i, ne_j) = AC(ne_i) = AC(ne_j)$. This entails $NWD(ne_i, ne_j)$=0, thereby, $Sim_{NWD}(ne_i, ne_j) = 1$.
- From the numerator of Eq. 3, the higher the proportion of the joint occurrence of the two named-entities $AC(ne_i, ne_j)$, the smaller is the normalized distance $NWD(ne_i, ne_j)$, and, in turn, the higher the similarity score $Sim_{NWD}(ne_i, ne_j)$. To ***investigate the detailed behaviour with respect to individual parameters***, let us denote by A the set of Wikipedia articles containing named-entity ne_i and B the set of Wikipedia articles containing named-entity ne_j, and let x be the cardinality of the intersection of sets A and B corresponding to the number of articles of joint occurrences of both named-entities. Assume without loss of generality that $|A| < |B|$, then Eq. 3 is equivalent to

$$NWD(ne_i, ne_j) = \frac{log_2|B| - log_2x}{log_2N - log_2|A|} \tag{9}$$

From the preceding, it is straightforward that:

- NWD is decreasing with respect to x.
- If x remains constant, then NWD is monotonically increasing with respect to the size of A as well as size of B, so, the similarity Sim_{NWD} is monotonically decreasing.
- If x remains constant while the size of both A and B increases in the same order of magnitude, then the normalized distance increases as well, which, in turn, induces a decrease of the similarity score. To see it, let us consider an increase of magnitude of y of each of A and B, then the difference with former normalized distance (without increase of A and B) is

$$\frac{log_2(|B| + y) - log_2x}{log_2N - log_2(|A| + y)} - \frac{log_2|B| - log_2x}{log_2N - log_2|A|}$$

The latter expression is positively valued because from the monotonicity of the logarithm function, it follows that $log_2N - log_2(|A| + y) < log_2N - log_2|A|$, and $log_2(|B| + y) - log_2x > log_2|B| - log_2x$. Furthermore, the above result is still valid even if the expansion of A and B is not uniform; namely, for $y, z > 0$, it holds that

$$\frac{log_2(|B| + y) - log_2x}{log_2N - log_2(|A| + z)} - \frac{log_2|B| - log_2x}{log_2N - log_2|A|} > 0 \tag{10}$$

The above shows that any expansion of the initial set of articles containing any of the named-entities while keeping the number of articles pertaining to joint occurrences constant induces an increase of the normalized distance, and therefore, a decrease of similarity score.

- Since the values of the cardinality in the logarithmic functions in Eq. 3 are integer valued, it turns out that the ***ranges of values*** of the normalized distance, and thereby of the similarity function are not equally distributed. Indeed, for $x = 1$, we have $NWD(ne_i, ne_j) = \frac{log_2|B|}{log_2N - log_2|A|}$.The latter is maximal when minimizing |A| and maximizing |B|; i.e., by choosing a pair of named-entities such that the first one has most number of entries while the second one has the least number of entries in Wikipedia. Also, given that the number N is of several orders of magnitude of any |A| or |B|, it holds that $NWD(ne_i, ne_j) < 1$. On the other hand, as soon as there are no co-occurrences $(x = 0)$, $NWD(ne_i, ne_j) \to \infty$. This makes all the range of values from 1 to ∞ ill-represented. This is mainly due to the absence of logarithm of numbers less than one in Eq. 3. Accordingly, the high value similarity scores are extensively dominant. This is especially important when deciding to assign a threshold value in order to trigger some decision related to the subsequent analysis based on the similarity score.

- A Special case of (5) corresponds to the situation where ***one sentence bears only a single named-entity while the second one bears many***. In this case, (5) can be rewritten as, assuming for instance NE_1 contains only ne_0.

$$Sim_{WP}(NE_1, NE_2) = \frac{1}{2}\left(\max_{ne_j \in NE_2} Sim_{NWD}(ne_0, ne_j) + \frac{\sum_{ne_j \in NE_2} Sim_{NWD}(ne_0, ne_j)}{|NE_2|}\right) \tag{11}$$

  Comparing Eq. 11 with the similarity of the pair of named-entities yielding the highest score turns out that the use of extra named-entities can either increase or decrease the individual similarity score depending on the contributions of other named-entities, since $Sim_{WP}(NE_1, NE_2) \geq \max_{ne_j \in NE_2} Sim_{NWD}(ne_0, ne_j)$ or $Sim_{WP}(NE_1, NE_2) \leq \max_{ne_j \in NE_2} Sim_{NWD}(ne_0, ne_j)$ are equally held. Nevertheless, trivially, the more the named-entities of NE_2 bear similarity with ne_0, the more the inequality $Sim_{WP}(NE_1, NE_2) \geq \max_{ne_j \in NE_2} Sim_{NWD}(ne_0, ne_j)$ is valid.
- Another interesting case of Eq. 5 relates to the existence of ***duplicated named-entities*** in either sentence of the pair. Namely, let us assume without loss of generality that the first sentence includes named-entities ne_0, ne_1 and ne_1, while sentence 2 includes named-entity ne_2, then Eq. 3 is equivalent to

$$Sim_{WP}(NE_1, NE_2) = \frac{1}{2}\left(\frac{Sim_{NWD}(ne_0, ne_2) + 2Sim_{NWD}(ne_1, ne_2)}{3} + max(Sim_{NWD}(ne_2, ne_0), Sim_{NWD}(ne_2, ne_1))\right)$$

  Comparing the latter with the result of similarity if duplication is omitted:

$$Sim_{WP}(NE_1, NE_2)^* = \frac{1}{2}\left(\frac{Sim_{NWD}(ne_0, ne_2) + Sim_{NWD}(ne_1, ne_2)}{2} + max(Sim_{NWD}(ne_2, ne_0), Sim_{NWD}(ne_2, ne_1))\right)$$

  reveals that

$$Sim_{WP}(NE_1, NE_2) - Sim_{WP}(NE_1, NE_2)^* = (Sim_{NWD}(ne_1, ne_2) - Sim_{NWD}(ne_0, ne_2))/6 \tag{12}$$

  Strictly speaking, the latter expression can either be positive or negative valued, which means that if the sentence contains duplicate named-entities, this will ultimately influence the overall similarity score. Nevertheless, it is also clear from Eq. 12 that if the duplicated named-entity bears more similarity to its counterpart in the pair sentence, then one can guarantee that the duplication would contribute positively to an increase of the overall similarity score.[4]

[4] Similar reasoning applies to the case of duplicated non-NEs.

References

Alguliev, R., & Aliguliyev, R. (2009). Evolutionary algorithm for extractive text summarization. *Intelligent Information Management*, *1*(02), 128.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning: Association for computational linguistics* (pp. 546–556).

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370–83.

Das, D., & Smith, N. A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Association for computational linguistics (Vol. 1, pp. 468–476).

Dias, G., Moraliyski, R., Cordeiro, J., Doucet, A., & Ahonen-Myka, H. (2010). Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Natural Language Engineering*, *16*(04), 439–67.

Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on computational linguistics: Association for computational linguistics* (p. 350).

Fernando, S., & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics* (pp. 45–52).

Finch, A., Hwang, Y.S., & Sumita, E. (2005). Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the third international workshop on paraphrasing (IWP2005)* (pp. 17–24).

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*, 443–498.

Gracia, J., Trillo, R., Espinoza, M., & Mena, E. (2006). Querying the web: A multiontology disambiguation method. In *Proceedings of the 6th international conference on web engineering*. ACM.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING* (pp. 466–471).

Habash, N., & Dorr, B. (2003). A categorial variation database for English. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology: Association for computational linguistics* (vol. 1, pp. 17–23).

Habib, M. B., & van Keulen, M. (2016). TwitterNEED: A hybrid approach for named-entity extraction and disambiguation for tweet. *Natural Language Engineering*, *22*(03), 423–56.

Hassan, S. (2011). *Measuring semantic relatedness using salient encyclopedic concepts*. Denton: University of North Texas.

He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1576–1586).

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. In *Transactions on knowledge discovery from data (TKDD)* (Vol. 2, p. 10). ACM.

Issa, F., Damonte, M., Cohen, S. B., Yan, X., & Chang, Y. (2018). Abstract meaning representation for paraphrase detection. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies* (Vol. 1 (Long Papers), pp. 442–452).

Ji, Y., & Eisenstein, J. (2013). Discriminative improvements to distributional sentence similarity. In *EMNLP* (pp. 891–896).

Kim, S. N., & Baldwin, T. (2013). A lexical semantic approach to interpreting and bracketing English noun compounds. *Natural Language Engineering*, *19*(03), 385–407.

Kozareva, Z., & Montoyo, A. (2006). Paraphrase identification on the basis of supervised machine learning techniques. In *Advances in natural language processing* (pp. 524–533). Berlin: Springer.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning* (pp. 957–966).

Liu, J., & Birnbaum, L. (2007). Measuring semantic similarity between named entities by searching the web directory. In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence: IEEE computer society* (pp. 461–465).

Madnani, N., Tetreault, J., & Chodorow, M. (2012). Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics on human language technologies: association for computational linguistics* (pp. 182–190).

Malik, R., Subramaniam, L. V., & Kaushik, S. (2007). Automatically selecting answer templates to respond to customer emails. In *IJCAI* (Vol. 7, pp. 1659–1664).

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI* (Vol. 6, pp. 775–780).

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Miller, G. A., & Hristea, F. (2006). WordNet nouns: Classes and instances. *Computational Linguistics*, *32*(1), 1–3.

Mohamed, M., & Oussalah, M. (2014). A comparative study of conversion aided methods for WordNet sentence textual similarity. In *COLING* (pp. 37–42).

Mohamed, M., & Oussalah, M. (2016). An iterative graph-based generic single and multi document summarization approach using semantic role labeling and wikipedia concepts. In *Proceedings of 2016 IEEE international conference on big data computing service and applications (BigDataService)* (pp. 117–120). IEEE.

Nothman, J., Curran, J. R., & Murphy, T. (2008). Transforming Wikipedia into named-entity training data. In *Proceedings of the Australian language technology workshop* (pp. 124–132).

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004: Association for computational linguistics* (pp. 38–41).

Ponzetto, S. P. (2010). *Knowledge acquisition from a collaboratively generated encyclopedia*. Amsterdam: IOS Press.

Qiu, L., Kan, M. Y., & Chua, T. S. (2006). Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 conference on empirical methods in natural language processing: Association for computational linguistics* (pp. 18–26).

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named-entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning: Association for computational linguistics* (pp. 147–155).

Roth, D., & Zelenko, D. (1998). Part-of-speech tagging using a network of linear separators. In *Proceedings of the 17th international conference on Computational linguistics: Association for computational linguistics* (Vol. 2, pp. 1136-1142).

Rus, V., McCarthy, P. M., Lintean, M. C., McNamara, D. S., & Graesser, A. C. (2008). Paraphrase identification with Lexico-syntactic graph subsumption. In *FLAIRS conference* (pp. 201–206).

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems* (pp. 801–809).

Summers, E., & Cassidy, B. (2011). *WWW::Wikipedia—Automated interface to the Wikipedia*. CPAN.

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems*, *50*, 260–78.

Wan, S., Dras, M., Dale, R., & Paris, C. (2006). Using dependency-based features to take the "para-farce" out of paraphrase. In *Proceedings of the Australasian language technology workshop* (Vol. 2006).

Wang, Z., Mi, H., & Ittycheriah, A. (2016). Sentence Similarity Learning by Lexical Decomposition and Composition. In *Proceedings of COLING 2016 conference on technical papers* (pp. 1340–1349).

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics: Association for computational linguistics* (pp. 133–138).