Semantic and Heuristic Based Approach for Paraphrase Identification

Muhidin A. Mohamed*
School of Engineering and Applied Sciences
Aston University, Birmingham B4 7ET UK
Email: m.mohamed10@aston.ac.uk

Mourad Oussalah
Faculty of Information Technology
University of Oulu, Finland
Email: mourad.oussalah@oulu.fi

***Abstract*—In this paper, we propose a semantic-based paraphrase identification approach. The core concept of this proposal is to identify paraphrases when sentences contain a set of named-entities and common words. The developed approach distinguishes 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 the crowd-sourced knowledge in Wikipedia database. Besides, we improve WordNet similarity measure by nominalizing verbs, adjectives and adverbs with the aid of Categorial Variation database (CatVar). The paraphrase identification system is then evaluated using two different datasets; namely, Microsoft Research Paraphrase Corpus (MSRPC) and TREC-9 Question Variants. Experimental results on the aforementioned datasets show that our system outperforms baselines in the paraphrase identification task.**

***Keywords*—Paraphrase identification; Sentence semantic similarity; Word category subsumption; named-entity relatedness.**

I. Introduction

Paraphrases are sentences conveying the same meaning using different wording. The identification of whether two sentences are paraphrases requires us to explicitly quantify the amount of semantic overlap between their textual expressions. This typically involves measuring the extent to which a pair of words, phrases or sentences are semantically related to each other using either large corpora, e.g., Wikipedia [1] or semantic features from knowledge networks such as WordNet [2]. Many of the existing paraphrase detection approaches are substantially built on WordNet taxonomy [3]. The latter is a lexical database where English words are grouped into sets of synsets and interlinked by means of conceptual-semantic and lexical relations [4]. Because of its hierarchical taxonomy, WordNet enables the construction of useful word and sentence level semantic similarity measures allowing the semantic overlap between paraphrases to be established and quantified.

Having said that, WordNet-based semantic similarity measures have a number of inherent limitations. For example, the taxonomic relations are only available for noun and verb categories which means that one can only compute the semantic similarity between a pair of nouns or verbs. This excludes other PoS categories, such as adverbs and adjectives, from the semantic similarity calculus. In addition, there is a strong discrepancy between the hierarchies of the noun and verb categories where the noun entity is more developed in terms of the hierarchical taxonomy and associated depth [5]. This renders the semantic similarity of the nouns and that of the verb entities somehow biased. Linked with the above, many of the commonly known named-entities are mostly absent from the WordNet lexical database [6], substantially undermining the ability of WordNet-based similarity measures to accurately capture the semantic overlap between texts.

In this paper we address the aforementioned problems by investigating how the incorporation of crowd-sourced knowledge in Wikipedia and semantic relations in WordNet improves sentence paraphrase identification. The WordNet-based similarity measure is severally enhanced by supplementing it with Categorial Variation database for the purpose of subsuming verb, adverb and adjective categories under derivationally related nouns in WordNet taxonomy. The main contributions of this paper are: the improvement of WordNet semantic-based paraphrase identification by converting all possible loosely encoded and non-hierarchized word categories (e.g., verbs, adverbs and adjectives) to their corresponding nouns using CatVar database. This allows us to cover a wide range of lexical items that would not have been matched without such conversion. Besides, 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. Next, we have applied a developed named entity semantic relatedness measure to the task of paraphrase identification using entity co-occurrences in Wikipedia articles. Then, the enhanced WordNet similarity and the Wikipedia-based named-entity semantic relatedness measures are integrated to form a combined paraphrase detection system. The proposed approach is finally evaluated using a set of publicly available datasets where a comparison with baselines 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 WordNet semantic similarity, and the use of PoS conversion through the aid of CatVar database. Section 4 copes with a metric introduced for measuring named-entity semantic relatedness using Wikipedia. Section 5 details our combined approach for computing the semantic similarity employing both Wikipedia and WordNet. Next, we provide some experimental results in Section 6 and draw conclusions in Section 7.

*Most of this work was done while a PhD student at the Uiversity of Birmingahm.

II. Related Works

Important research has been conducted to identify short paraphrases using different strategies. Paraphrase detection methods can be broadly categorized into three high level classes on the basis of their information source, namely; corpus-based, knowledge-based and hybrid methods. First, the use of strategies entirely or substantially based on corpus statistics provided some success in the paraphrase identification (PI) problem [7]–[9]. Ji and Eisenstein [8] used a simple distributional similarity model by designing a discriminative term-weighting metric called TF-KLD while indicating that their new metric outperforms the widely used TF-IDF weighing scheme. In addition, Blacoe & Lapata [7] employed three distributional representations of text; simple semantic space, sytax-aware space and word embeddings. Alternatively, Madnani & Chodorow [10] investigated the feasibility of machine translation approaches with WordNet for paraphrase detection.

On the other hand, one acknowledges the work of Fernando and Stevenson [11], who used word level similarities derived from WordNet taxonomy. Similarly, Das and Smith [12] utilized quasi-synchronous dependency grammars in a probabilistic model incorporating WordNet. Furthermore, the work of Kozareva and Montoyo [13] 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 [14] suggested a new approach called Salient Semantic Analysis (SSA) using context meaning from Wikipedia links.

Related to the previous, some researchers used hybrid approaches which combine different techniques. For instance, in [15], authors 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, other researchers suggested a two-phase framework that makes paraphrase identification judgment rely on the dissimilarity between sentences [16]. In a more entailment oriented approach, Rus et al. [17] built a graphical representation of text by mapping relations within its syntactic dependency trees. The researchers used synonymy and antonymy relations from WordNet to measure word overlap and to handle Text-Hypothesis negation in textual entailment. Pairwise semantic features of single words and multiword expressions from syntactic trees have also been utilized in [18]. They made use of syntactic parse trees, corpus based training and feature learning. Moreover, Neural networks have been recently gaining research interest in the area of paraphrase identification [19].

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 determine the existence of a paraphrase relationship between sentences. Similar to [11]–[13], [15], [17], 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 [20]. Second, inspired by Normalized Google Distance (NGD) algorithm [21], 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 [22].

III. Paraphrase Identification Based on WordNet Semantic Relations

Prior to word-similarity computation, sentence texts are processed using standard natural language processing packages and parsers, such as the Illinois PoS and Named-entity Taggers [23], [24] in order to identify the various tokens, their PoS category and the presence of named-entities. The latter may sometimes be 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.

A. Traditional WordNet Similarity

WordNet is a hierarchical lexical database for English developed at Princeton University [4]. 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. For the WordNet-based word-to-word similarity and relatedness, we used the implementation described in [25]. Here, we considered the common Wu & Palmer measure [26], which relies solely on the path lengths between WordNet concepts.

With the traditional 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 and 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 [15] 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 Equation 1. 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. The function $Sim(w, x)$ in Equation (1) represents the similarity between the two words, w and x, while $|S_A|$ (resp. $|S_B|$) stands for the number of words in S_A (resp. S_B).

$$Sim(S_A, S_B) = \frac{1}{2}\left[\frac{\sum_{w \in S_A} \max_{x \in S_B} Sim(w,x)}{|S_A|} + \frac{\sum_{w \in S_B} \max_{x \in S_A} Sim(w,x)}{|S_B|}\right] \quad (1)$$

B. WordNet Similarity with Part of Speech Conversion

As highlighted in equation (1), the traditional approach of WordNet semantic similarity is derived from average over all one-to-one word level similarities of the two sentences in comparison. Nevertheless, the above average is restricted to pairs of words that belong either to verb or noun PoS categories only. This leaves other important sentence tokens, such as proper nouns, adverbs and adjectives unaccounted for, resulting in the failure of properly utilizing WordNet graph connectivity. A block diagram of the proposed CatVar-aided sentence textual similarity measure is depicted in Fig. 1. It 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. As shown in the figure, 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. To exemplify our reasoning, consider the pair of semantically identical sentences in Example 1 with different wording.

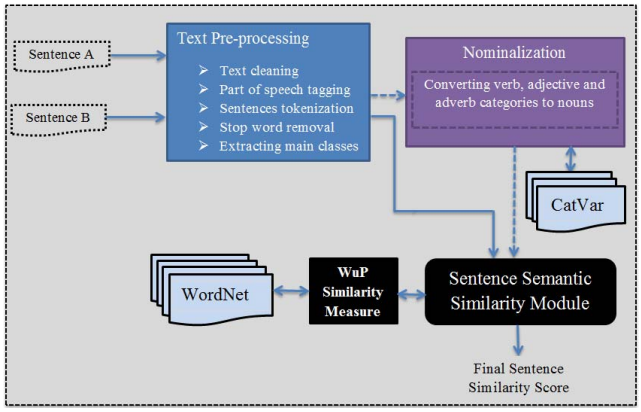

Fig. 1: Sentence semantic similarity assisted with PoS conversion.

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.

Basic text pre-processing tasks including tokenization, normalization, and stop-words removal reduce the sentences to their content words with S_1 yielding *(transformation, word, forms, improvement, sentence, similarity)* and *(converting, word, forms, enhances, sentence, similarity)* for S_2. It is therefore 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 (which will be explained shortly) takes place , *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 Equation 1 to the nominalised sentences increased 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. To this end, three primary word classes; namely, verbs, adjectives and adverbs are transformed to their equivalent nouns using Categorial Variation database (CatVar) as summarized in Algorithm 1. This gives the opportunity of overcoming the stated part of speech boundary limitation of WordNet.

Definition 1. *Let $T = \{w_1, w_2, w_n\}$ be a bag-of-words of sentence where words w_i $(i = 1, n)$ do not necessarily belong to same part of speech. The word category conversion is a mapping function $f : T \rightarrow T'$, where $T' = \{t_1, t_2, ..., t_n\}$ and every non-noun primary word w_i in it is mapped to its most equivalent noun t_i using CatVar database.*

Algorithm 1 Word Category Conversion using CatVar

WCConversion $(S, targetCategory)$
$W \leftarrow tokenize(S)$
$\bar{W} \leftarrow \{\}$
$Open(CatVarDB)$
for all $w_i \in W$ **do**
  if $w_i \in InflectedWords$ **then**
    $PoS_{w_i} \leftarrow ExtractPoSTag(w_i)$
    $VFS \leftarrow ValidForms(w_i)$
    for $w_j \in VFS$ **do**
      $PoS_{w_j} \leftarrow ExtractPoSTag(w_j)$
      if $PoS_{w_j} \equiv PoS_{w_i}$ **then**
        $w_i \leftarrow w_j$
        last;
      end if
    end for
  end if
  $CurrentCluster \leftarrow FirstCluster(CatVarDB)$
  while $CurrentCluster \neq EOF$ **do**
    if $w_i \in CurrentCluster$ **then**
      $cw \leftarrow convert(w_i)$
      last;
    end if
  end while
  $\bar{W} \leftarrow \bar{W} \cup \{cw\}$
end for
$Return\ \bar{W}$

C. CatVar-Assisted PoS Conversion Algorithm

Categorial Variation Database (CatVar) is a lexical resource of morphological derivations for English words sharing a common stem, e.g., $research_V, researcher_N, researchable_{AJ}$ [20]. The PoS conversion augmented with CatVar, summarized in Algorithm 1, is a simple process. It 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. The 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.
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.

IV. Wikipedia-based Named-entity semantic relatedness

The word named-entity as used today in text mining and Natural Language Processing (NLP) was introduced in the Sixth Message Understanding Conference [13]. 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 a 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 names 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 [27]. This study uses Wikipedia utility for named-entity similarity approximation underpinned with the NGD algorithm.

In English and other languages, some words have a high 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. For example, the number of Wikipedia articles containing the names *FIFA* and *Joseph S Blatter* singly were 33123 and 291 respectively while the Wikipedia pages in which the named-entities occurred jointly were 267, yielding intuitively high similarity score between the two concepts. Since its foundation in 2001, Wikipedia has grown in both popularity and size leading to an increased usage among the NLP research community. The encyclopaedia contains over 32 million articles in 260 languages where its English version had more than 5.5 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 NGD on the Wikipedia database for the task of the named-entity semantic similarity quantification are summarized below:

1) Empirical and survey research found that around 74% of Wikipedia pages describe named-entities [22], 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.

Sometimes, a given name may 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 remains 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.

As previously indicated, our current approach for named-entity semantic relatedness is based on entity co-occurrence in the form of Wikipedia article counts underpinned by the NGD, a mathematical theory based on Information Distance and Kolmogorov Complexity [21]. Especially, we downscaled NGD to Wikipedia. In other words, if e_i and e_j are two entities, we extract the number of Wikipedia articles $A(e_i)$, $A(e_j)$, & $A(e_i, e_j)$ for the entities e_i, e_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, $NWD(e_i, e_j)$, can be computed as:

$$NWD(e_i, e_j) = \frac{\max[log_2 A(e_i), log_2 A(e_j)] - log_2 A(e_i, e_j)}{log_2 N - \min[log_2 A(e_i), log_2 A(e_j)]} \quad (2)$$

The parameter N in the denominator is the total number of English Wikipedia articles. Next, the similarity between named-entities e_i and e_j is computed using an exponential function that would guarantee the score to be normalized in

the unit interval:

$$Sim_{NWD}(e_i, e_j) = e^{-NWD(e_i,e_j)} \quad (3)$$

From an implementation perspective, Equation 3 turns out to be a quite simple, effective and language independent named-entity similarity measure. 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 pursed in this paper. To appreciate the measure, consider the pair of named entities *IEEE* and *FIFA* with the following Wikipedia article counts retrieved from the encyclopaedia: $A(IEEE) = 13225, A(FIFA) = 46,218, A(IEEE, FIFA) = 30, N = 4738956$. The application of expression (3) to the semantic relatedness between the two names yields a score of $e^{-1.248} = 0.28707$, which is an intuitive answer for such entities with low co-occurrence probability. Typically, a sentence text may contain more than one named entity; therefore, expression (3) is extended to determine the sentence-to-sentence semantic similarity in view of their named-entities only. Let us assume that E_A represents the set of named-entities contained in S_A and E_B the set of named-entities in S_B. Then, their associated Wikipedia-derived similarity is calculated as:

$$Sim_{WP}(E_A, E_B) = \frac{1}{2}\left[\frac{\sum_{e_i \in E_A} \max_{e_j \in E_B} Sim_{NWD}(e_i, e_j)}{|E_A|} + \frac{\sum_{e_j \in E_B} \max_{e_i \in E_A} Sim_{NWD}(e_i, e_j)}{|E_B|}\right] \quad (4)$$

V. INTEGRATING CONVERSION AIDED WORDNET AND WIKIPEDIA FOR SENTENCE SEMANTIC SIMILARITY

Fig. 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[1], 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 severally and jointly. The joint counts, which are used in Equation (2), 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 [25] for computing conceptual relatedness of individual words after applying the CatVar-aided PoS conversion. In addition to the traditional 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 identifying and labelling all proper nouns in the text. On the other hand, token classification is a post tagging step in which sentence tokens are split into content word and named-entity vectors.

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

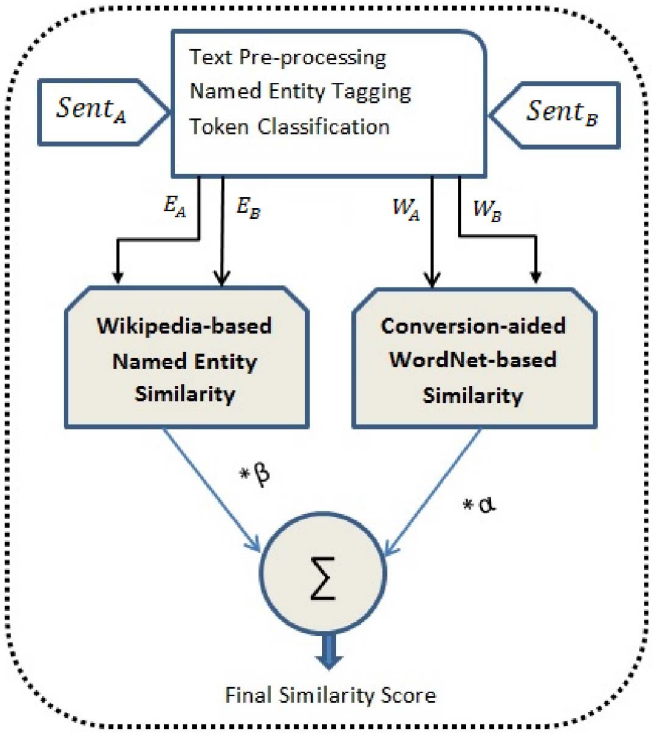

Fig. 2: Combined conversion-aided WordNet and Wikipedia similarity measures.

In Fig. 2, the inputs to the WordNet-based subsystem denoted by the notations W_A and W_B correspond to the non-named entity word vectors of the corresponding sentences S_A and S_B, respectively. A generic formula for the conversion-aided WordNet-based semantic similarity between these non-named-entity sets yields:

$$Sim_{WN}(W_A, W_B) = \frac{1}{2}\left[\frac{\sum_{w_i \in W_A} \max_{w_j \in W_B} Sim(w_i, w_j)}{|W_A|} + \frac{\sum_{w_j \in W_B} \max_{w_i \in W_A} Sim(w_i, w_j)}{|W_B|}\right] \quad (5)$$

Finally, the overall semantic similarity between any two sentences in comparison, accounting for the occurrence of content words and named-entities, is given as the convex combination of the Sim_{WN} and Sim_{WP}:

$$Sim(S_A, S_B) = \alpha Sim_{WN}(W_A, W_B) + \beta Sim_{WP}(E_A, E_B) \quad (6)$$

Where 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. We used a simple modelling approach based on the number of entity and word tokens to quantify the coefficients. Using the terminology in Fig. 2, α, β are formulated as:

$$\alpha = \frac{|W_A| + |W_B|}{|W_A| + |W_B| + |E_A| + |E_B|} \quad (7a)$$

$$\beta = \frac{|E_A| + |E_B|}{|W_A| + |W_B| + |E_A| + |E_B|} \quad (7b)$$

This follows the statistical argumentation that the more the number of tokens associated to WordNet is higher than the number of named-entities in a sentence, the more one expects the contribution of Sim_{WN} to be of larger significance than that of Sim_{WP} in the integrated model. The use of word proportions from the sentence pairs in Equation 7 as coefficients for the combination of the two similarity components (Equation 6) has some desirable attributes. First, it conforms with unity sum. Second, it serves as a weighting control strategy for the relative contribution of each similarity component. For instance, in the boundary case of Equation 7, it is easy to see that if there are no named-entities in the pair of sentences, then $|E_A| = |E_B| = 0$, which entails $\alpha = 1 \,\&\, \beta = 0$, so that $Sim(S_A, S_B) = Sim_{WN}(W_A, W_B)$. Similarly, if the pair of sentences are primarily constituted of named-entities, then $\beta = 1 \,\&\, \alpha = 0$ which entails $Sim(S_A, S_B) = Sim_{WP}(E_A, E_B)$. Strictly speaking, even in the case where only one sentence contains named-entity (resp. non-named-entity token), it holds that $|E_A| = |E_B| = 0$ (resp. $|W_A| = |W_B| = 0$) as the Wikipedia-based similarity can only be performed if entities in both sentences possess entries in the Wikipedia database (resp. existence of noun counterpart in the other sentence).

A. An Illustrative Example

For exemplification, consider Examples 2 which highlights the functioning of the overall hybrid approach. At the same time, it sheds light on the advantages of the hybrid approach with respect to either individual WordNet-based or Wikipedia-based similarity.

Example 2:
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. Table I presents a comparison of final similarity scores after applying traditional WordNet (Section III-A), WordNet with CatVar conversion (Section III-B) and the proposed hybrid method (Section V). From Table I, 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. A nominalization (changing verbs to nouns - *founded* only in this case) is incorporated in the case of the CatVar-aided measure raising the sentence similarity score to *0.19*. In addition to applying word PoS conversion, Wikipedia-based named-entity similarity (Section IV) is augmented to form the hybrid method as given in Table I. 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.

VI. EXPERIMENTS

In this section we report the experiments we conducted to test and evaluate the proposed paraphrase identification approach and the results we acquired.

TABLE I Comparison of Different Similarity measures using the sentences in Example 2

Applied Similarity Scheme	Final Similarity score
Traditional WordNet similarity	0*
CatVar-aided WordNet Similarity	0.19
Hybrid method	0.76

A. Evaluation Datasets

We conducted evaluation experiments on two datasets, namely, Microsoft Research Paraphrase Corpus (MSRPC) and TREC-9 Question Variants. MSRPC is a human annotated dataset created from news articles on the web for the evaluation of machine-based paraphrase identification tasks [28]. Its creation has undergone a series of refining stages from which developers finally produced a set of 5801 sentence pairs. We used 750 sentence pairs extracted from the training data to determine an optimum demarcation threshold for the classification of sentence pairs as positive or negative paraphrases. For the performance evaluation, we used the entire test data (1725 pairs). Similar to the MSRPC, the TREC-9 Question Variants [2] is created by human assessors to describe semantically identical but syntactically different questions. The dataset contains 54 sets with 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 pairs of sentences from the dataset classified into semantically equivalent, and dissimilar questions.

B. Performance Metrics

Our similarity based paraphrase identification approach produces four possible outcomes. In the first case, two semantically equivalent sentences might be identified as positive paraphrases of one another, commonly referred to as true positive (TP). Secondly, a false negative (FN) occurs when a pair is incorrectly classified as non-paraphrases. Thirdly, there exists a situation known as false positive (FP) where a given sentence pair is semantically non-equivalent, but the system labels them as paraphrases. Lastly, when a semantically unrelated sentence pair is correctly predicted as non-paraphrases, it is referred to as true negative (TN). The performance of the hybrid method is evaluated using four different metrics (Accuracy, Precision, Recall, and F-measure).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

Equation (8) indicates the proportion of the correct prediction (either as paraphrases or non-paraphrases).

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

In this context, the precision (9) is the proportion of real paraphrases over the total pairs identified as semantic equivalents.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

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

Unlike the precision, recall (10) measures the proportion of paraphrases which has been correctly classified.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

Empirical evidences have shown the existence of a trade-off between precision and recall. Consequently, the F-measure (11) has been developed as a compromise and a proper measure that combines the effect of the two metrics.

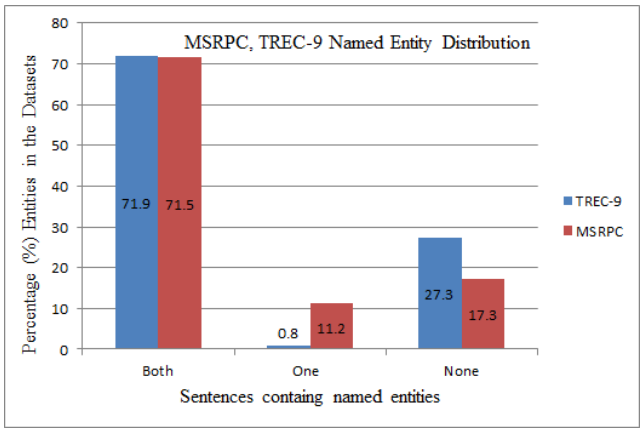

Fig. 3: Named-entity distribution in the TREC-9 and MSRPC datasets; Both: both sentences of the pair contain named-entities; One: only one sentence of the pair has named-entities; None: None of the sentence pair bears named-entities.

C. Results and Discussion

Firstly, we ran a set of training experiments using 750 sentence pairs from MSRPC and 30% of the total TREC-9 dataset while reserving the remaining 70% and the entire MSRPC testing data (1725 pairs) for testing and evaluation experiments. During this training, we determined a threshold value of 0.7 to be the optimum demarcation criteria. 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.

Fig. 3 shows that more than 71% of the paraphrase pairs contain one or more named-entities in both the TREC-9 and MSRPC datasets. This highlights the importance of these textual components often underestimated in the state-of-the-art knowledge-based similarity approaches. 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*.

TABLE II System notations

$CosSim$	Cosine similarity
$WNwoC$	WordNet without conversion
$WNwCC$	WordNet with CatVar conversion
$NeSim$	Wikipedia-based entity similarity
Hm	Proposed hybrid PI approach

The primary focus of our experiments is on the evaluation of the hybrid method. However, prior to the combined method (Hm), we performed a rather superfluous 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. It is also to use them as baselines for comparison. Moreover, we selected two other similarity measures; namely, cosine ($CosSim$) and conventional WordNet ($WNwoC$) as additional baseline comparators. 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 Section III-A. These two benchmark methods are evaluated against our proposed conversion-aided WordNet, the Wikipedia-based and the hybrid methods. Table III and Table IV chart the system-baseline comparison for TREC-9 and MSRPC datasets respectively, while related notations are defined in Table II. Notably, the system's better performance on the TREC-9 dataset, as in Table III, might be due to either the dominance of named-entities after the elimination of stop words, and/or its smaller size and short sentence lengths as compared to the MSRPC corpus. What is very interesting in the findings, though, is the fact that 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

TABLE III System-baseline comparison on the TREC-9 dataset - all figures rounded up to 3 SF

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**	**0.897**	**0.871**

TABLE IV System-baseline comparison on the MSRPC dataset - all figures rounded up to 3 SF

Measure	Precision	Recall	F-measure	Accuracy
$WNwoC$	0.826	0.559	0.667	0.558
$CosSim$	**0.907**	0.314	0.466	0.432
$WNwCC$	0.819	0.802	0.810	0.703
$NeSim$	0.794	0.559	0.656	0.537
Hm	0.820	**0.887**	**0.852**	**0.757**

semantic extraction. Therefore, it is not surprising for the combined approach to show better performance in comparison to the separate sub-systems. From the experimental results, it is apparent that both the CatVar-aided WordNet scheme and the hybrid method attained a significant improvement over the baselines.

Overall, from Tables III-IV, it is evident that the combination of Wikipedia and WordNet has clearly improved the paraphrase identification performance, where the proposed hybrid system outperforms baselines. 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.

VII. Conclusion

We described an integrated sentence paraphrase identification system. 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, namely 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. Scores were 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. The proposal has been applied to two publicly available datasets, namely, the MSRPC and TREC-9. Obtained experimental results show that our system outperforms baselines.

References

[1] M. Mohamed and M. Oussalah, "An iterative graph-based generic single and multi document summarization approach using semantic role labeling and wikipedia concepts," in *IEEE International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2016, pp. 117–120.

[2] M. A. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, and L. Villaseñor-Pineda, "Semantically-informed distance and similarity measures for paraphrase plagiarism identification," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–8, 2018.

[3] S. Kim and T. Baldwin, "A lexical semantic approach to interpreting and bracketing english noun compounds," *Natural Language Engineering*, vol. 19, pp. 385–407, 2013.

[4] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[5] G. Miller and F. Hristea, "Wordnet nouns: Classes and instances," *Computational Linguistics*, vol. 32, pp. 1–3, 2006.

[6] S. Ponzetto, "Knowledge acquisition from a collaboratively generated encyclopedia," *IOS Press*, 2010.

[7] W. Blacoe and M. Lapata, "A comparison of vector-based representations for semantic composition," in *Proceedings of EMNLP*, 2012, pp. 546–556.

[8] Y. Ji and J. Eisenstein, "Discriminative improvements to distributional sentence similarity," in *Proceedings of EMNLP*, 2013, pp. 891–896.

[9] A. Eyecioglu and B. Keller, "Knowledge-lean paraphrase identification using character-based features," in *Conference on Artificial Intelligence and Natural Language*. Springer, 2017, pp. 257–276.

[10] T. J. Madnani, N. and M. Chodorow, "Re-examining machine translation metrics for paraphrase identification," in *Proceedings of NAACL*, 2012, pp. 182–190.

[11] S. Fernando and M. Stevenson, "A semantic similarity approach to paraphrase detection," in *Proceedings of 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 2008, pp. 45–52.

[12] D. Das and N. Smith, "Paraphrase identification as probabilistic quasi-synchronous recognition," in *Proceedings of the 47th Annual Meeting of the ACL*, 2009, pp. 468–467.

[13] Z. Kozareva and A. Montoyo, "Paraphrase identification on the basis of supervised machine learning techniques," in *Proceedings of Advances in natural language processing*, 2006, pp. 524–533.

[14] S. Hassan, "Measuring semantic relatedness using salient encyclopedic concepts," Ph.D. dissertation, University of North Texas Denton, 2011.

[15] C. C. Mihalcea, R. and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity."

[16] K. M. Qiu, L. and T. Chua, "Paraphrase recognition via dissimilarity significance classification," in *Proceedings of EMNLP*, 2006, pp. 18–26.

[17] M. P. L. M. M. D. Rus, V. and A. Graesser, "Paraphrase identification with lexico-syntactic graph subsumption," in *Proceedings of FLAIRS*, 2008, pp. 201–206.

[18] H. E. P. J. M. C. Socher, R. and A. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proceedings of Advances in Neural Information Processing Systems*, 2011, pp. 801–809.

[19] G. K. He, H. and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Proceedings of EMNLP*, 2015, pp. 1576–1586.

[20] N. Habash and B. Dorr, "A categorial variation database for english," in *Proceedings of NAACL*. Association for Computational Linguistics, 2003, pp. 17–23.

[21] R. Cilibrasi and P. Vitanyi, "The google similarity distance," *IEEE Transactions on knowledge and data engineering*, pp. 370–383, 2007.

[22] C. J. Nothman, J. and T. Murphy, "Transforming wikipedia into named entity training data."

[23] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009, pp. 147–155.

[24] D. Roth and D. Zelenko, "Part of speech tagging using a network of linear separators," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*. Association for Computational Linguistics, 1998, pp. 1136–1142.

[25] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet:: Similarity: measuring the relatedness of concepts," in *Demonstration papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 38–41.

[26] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[27] M. B. Habib and M. Van Keulen, "Twitterneed: A hybrid approach for named entity extraction and disambiguation for tweet," *Natural language engineering*, vol. 22, no. 3, pp. 423–456, 2016.

[28] Q. C. Dolan, B. and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *Proceedings of ACL*. ACL.