On Web based Sentence Similarity for Paraphrasing Detection

Mourad Oussalah and Panos Kostakos
Center for Ubiquitous Computing, University of Oulu, P.O.Box 4500, FIN-90014, Oulu, Finland

Keywords: Semantic Similarity, Text Mining, Paraphrasing.

Abstract: Semantic similarity measures play vital roles in information retrieval, natural language processing and paraphrasing detection. With the growing plagiarisms cases in both commercial and research community, designing efficient tools and approaches for paraphrasing detection becomes crucial. This paper contrasts web-based approach related to analysis of snippets of the search engine with WordNet based measure. Several refinements of the web-based approach will be investigated and compared. Evaluations of the approaches with respect to Microsoft paraphrasing dataset will be performed and discussed.

1 INTRODUCTION

Paraphrase detection is found to be critical in information extraction, information retrieval, summarization, question-answering, plagiarism identification, among others. Nevertheless, paraphrase detection is acknowledged as a challenging natural language processing task because of inherent difficulty in grasping the meaning of individual phrases. Although standard approaches for this task relies intensively on purely lexical based matching through counting the number of matching tokens of the two-phrases (Zhang and Patrick, 2005), other approaches which make use of semantic similarity features with/without other heuristics begin to emerge. For instance, Dolan and Brockett (2004; 2005) used string edit distance and a heuristic that pairs sentences from different stories in the same cluster. Islam and Inkpen (2007) used a modified version of the longest common subsequence string matching algorithm. Mihalcea et al. (2006) used corpus-based and knowledge-based measures of similarity. Fernando and Stevenson (2008) proposed a similarity matrix approach that makes use of WordNet based semantic similarity (Fellbaum, 1998). Collobert and Weston (2008) advocated a deep neural network based approach that uses word feature representations.

Motivated by Bollegala et al. 's (2007) work on web-based word similarity that exploits the outcome of web-search engine, this paper proposes a new paraphrasing detection method that exploits both the page count and the semantic redundancy of the snippets outputted by each of the two queries (phrases) to be compared. Intuitively, one expects that similar queries would yield similar outcomes, in terms of web search outcome. The latter includes both the number of outputs (page count) and the content of each link (including the snippet expression). However, given the complexity of the search engine operation and the subjectivity that may pervade the meaning of the query, such intuition is rarely fully verified, which motivates further reasoning in order to narrow this gap. For this purpose, several approaches will be examined and contrasted. Evaluations using Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) will be carried out. Section 2 of this paper presents the background and related work. Section 3 emphasizes our suggested web-based paraphrase detection approach. Testing using MRPC dataset is presented in Section 4. Finally, conclusion and perspective work are highlighted in Section 5.

2 BACKROUND AND RELATED WORK

The availability of word lexical database, e.g., WordNet (Fellbaum, 1998), where words are organized into synsets which are then encoded with conceptual and lexical IS-A relations enables creation of semantic distances among any word-pair (verb and noun categories).

Here we considered the commonly employed Wu and Palmer measure (1994), which is solely based on

Oussalah M. and Kostakos P.
On Web Based Sentence Similarity for Paraphrasing Detection.
DOI: 10.5220/0006581002890292
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 289-292
ISBN: 978-989-758-271-4

path length between WordNet concepts because of its simplicity and desire to omit corpus based effect. Mihalcea et al. (2006) proposed a canonical extension of the word-to-word semantic similarity to sentence-to-sentence similarity by averaging the maximum pairwise conceptual scores. More specifically, given two query sentences Q and S, the sentence similarity reads as:

$$\frac{Sim(Q,S)}{2} = \frac{\sum_{w\in Q;\, x\in S, POS(w)=POS(x)} \max sim(x,w)}{|S|} + \frac{\sum_{w\in Q;\, x\in S, POS(w)=POS(x)} \max sim(x,w)}{|Q|} \quad (1)$$

Where Sim(x,w) stands for Wu and Palmer WordNet conceptual similarity measure between word "x" of sentence S and a word "w" of sentence Q that has the same part of speech (POS) as P (only nouns and verb categories have their word-to-word similarity available). |Q| (resp. |S|) stands for the number of noun and verb tokens in the query sentence Q (resp. S).

Nevertheless, given the PoS tagger uncertainty where the prediction of the correct category is far to be fully accurate, a cautious counterpart of (1) examines all pairwise comparisons as follows:

$$\frac{Sim(Q,S)}{2} = \frac{\sum_{w\in Q;} \max_{x\in S} sim(x,w)}{|S|} + \frac{\sum_{w\in Q;} \max_{x\in S} sim(x,w)}{|Q|} \quad (2)$$

Alternatively, several researchers analyzed the word semantic similarity by evaluating the outcome of the search engine results. Cilibrasi and Vitanyi (2007) proposed the well-known normalized Google distance using only page counts for individual queries and joint occurrence. Bolegala et al. (2007) proposed WebJacquard coefficient that reads as, where H(.) denotes the number of hits of (.), and δ some predefined threshold.

$$WebJ(Q,S) = \begin{cases} 0 & \text{if } H(Q\cap S)\le\delta \\ \dfrac{H(Q\cap S)}{H(Q)+H(S)-H(Q\cap S)}, & \text{otherwise} \end{cases} \quad (3)$$

Sahami and Heilman (2006) quantified the semantic similarity between two queries using a TF-IDF (term frequency x inverse document frequency) model of snippets outputted by the search engine that accounts for contextual information. Chen et al. (2006) proposed the co-occurrence double checking (CODC) measure that counts the occurrences of word Q (resp. S) in snippets of word S (resp. P).

Bollegala et al. (2007) proposed an optimal combination of page counts-based co-occurrences measure and lexical patterns extracted from text snippets that uses SVM (support vector machine) classification. Strictly speaking, there are several limitations when attempting to extrapolate the word web based similarity to sentence-to-sentence similarity. First, the use of joint query (Q AND S) in the search engine often yields void result. Second, the complexity of the search operation which accounts for several other parameters (e.g. number of links, date, named-entities, context) renders the probability of having snippets which contain redundant wording rather very low. Third, inputting a phrase like expression to search engine involves several other lexical and semantic processing that goes beyond simple bag-of-word like reasoning. This motivates the approach put forward in the next section.

3 METHOD

3.1 Outline

Given two queries P and Q, let S(P) and S(Q) be the top-ranked set of snippets outputted by the search engine (for limitation of public search API, one shall only consider the first n snippets of each individual query. More formally, we have

$\mathbf{S}(P) = \{SP_1, SP_2, \ldots, SP_n\}$

$\mathbf{S}(Q) = \{SQ_1, SQ_2, \ldots, SQ_n\}$

where SP_i (resp. SQ_i) stands for the i^{th} tokenized snippet generated by query P (resp. Q), after filtering out stop words, symbols/characters. Individual tokens can also stand for composed words, if entry is found in WordNet lexical database.

We next compute two types of similarity measures among snippets. The first one builds on Chen et al. (2006) double checking model and the optimistic view of similarity.

$$S_{CODC}(SP_i, SQ_j) = \max\left(\frac{|SP_i \cap Q|}{|Q|}, \frac{|SQ_j \cap P|}{|P|} \right) \quad (4)$$

Especially, $S_{CODC}(SP_i, SQ_j)$ reaches its maximum value 1 whenever all tokens of P (resp. Q) are found in snippet SQ_j (resp. SP_i). The second similarity is a refinement of Fernando and Stevenson (2008) measure as:

$$S_{FSW}(SP_i, SQ_j) = \frac{\overrightarrow{SP_i} W \overrightarrow{SQ_j}}{|\overrightarrow{SP_i}||\overrightarrow{SP_j}|} \quad (5)$$

where $\overrightarrow{SP_i}$ ($\overrightarrow{SQ_j}$) corresponds to the binary vector of snippet SP_i (resp. SQ_j) with respect to a dictionary constituted of $SP_i \cup SQ_j$. In contrast to (Fernando and Stevenson, 2008), the matrix W carries pairwise similarity values calculated using both Wu & Palmer's semantic similarity and Wikipedia based measure such that

$$W_{k,l} = \max\left(Sim_{WP}(t_k,t_l), \frac{\max(\ln H(t_k), \ln H(t_l)) - \ln H(t_k,t_l)}{\ln(H(t_k)+H(t_l)) - \min(\ln H(t_k), \ln H(t_l))} \right) \quad (6)$$

where $H(t_k)$ represents the number of documents in Wikipedia containing token t_k. In other words, semantic similarity between two tokens corresponds to the best evaluation among Wu & Palma and Wikipedia based evaluation quantifications. The rationale behind the use of Wikipedia evaluation is to deal with situations in which individual semantic similarity returns void results. For instance, tokens "Ronald Reagan" and "US President" would yield zero using Wu & Palmer and 0.44 using Wikipedia normalized distance.

(6) corresponds to a canonical extension of Fernando and Stevenson (2008) sentence-similarity suggested for paraphrasing purpose, where its initial formulating is shown to outperform the tf-idf vector based representation as well as quantification (1) (Fernando and Stevenson, 2008).

Finally, the overall semantic similarity between the two queries P and Q is quantified as the best matching among the refined double checking model and weighted construction in the sense of (6); that is,

$$S(P,Q) = \max\left(\max_{i,j} S_{CODC}(SP_i, SQ_j), \max_{i,j} S_{FSW}(SP_i, SQ_j) \right) \quad (7)$$

In other words, the web-based semantic similarity between query P and Q is quantified according to the best evaluation between the extent of overlapping of one query into the top snippets of the second query, and the best pairwise evaluations of snippets associated to the two queries according to WordNet – Wikipedia evaluation metric similarity. It is noteworthy that expression (8) corresponds to a cautious attitude towards paraphrasing detection where a potential hit found in one snippet can yield a full similarity score. However, such reasoning seems in agreement with the prudent attitude of plagiarism detection where the officer is interested to candidates that will be subject to further checking, which would intuitively decrease the amount of false negatives in the system. A graphical summary of the approach is highlighted in Figure 1.

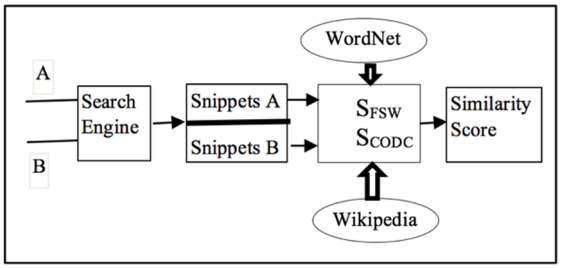

Figure 1: Outline of proposed method.

3.2 Exemplification

Consider the following two sentences in the Microsoft research paraphrasing corpus

A: "Amrozi accused his brother, whom he called 'the witness', of deliberately distorting his evidence."

B: "Referring to him as only 'the witness', Amrozi accused his brother of deliberately distorting his evidence."

The use of evaluation (1) yields *Sim(A , B)* = 0.81 while the quantification (8) yields *Sim(A , B)* = 1, which demonstrates a better agreement with human judgement.

Similarly, the following less trivial paraphrasing caseof the dataset:

C: "The former wife of rapper Eminem has been electronically tagged after missing two court appearances."

D: "After missing two court appearances in a cocaine possession case, Eminem's ex-wife has been placed under electronic house arrest."

yields *Sim(C, D)* = 0.63 when using (1) while the quantification (8) yields *Sim(C , D)* = 0.87, which again reinforces the aforementioned usefulness.

4 RESULTS AND DISCUSSION

We used the Microsoft Research Paraphrase Corpus, which is a standard resource for paraphrase detection task. It consists of 5801 sentence pair selected from Web news sources, which are hand labeled by human judges as whether the pair stands for a paraphrase or not. Table 1 present the average scores in terms of similarity measures for paraphrase and non-paraphrase cases. Results were also compared to evaluation based on (2), tf-idf cosine similarity, Fernando and Stevenson (2008) (implemented in the same spirit as (8)).

Next, one can use the available training dataset of MRPC in order to learn the threshold beyond which a similarity score is considered as a paraphrasing (a simple logistic regression were employed for this

purpose). This will be used to evaluate the overall accuracy of the approach. Overall results of the classification on testing MRPC dataset are highlighted in Table 2.

Table 1: Average similarity score for paraphrase and non-paraphrase cases in MRPC dataset

Method	Paraphrase	Non-paraphrase
Our method	88%	46%
Quantification (2)	64%	33%
WebJacquard	57%	41%
If-Idf Cosine Sim	42%	24%.
Method [5]	57%	36%

Table 2: Overall classification accuracy on MRPC testing dataset.

Method	Accuracy rate
Our method	84%
Quantification (2)	64%
WebJacquard	53%
If-Idf Cosine Sim	58%
Method [5]	71%

Results highlighted in Table 1 and Table 2 testify of the usefulness of the proposed approach that fruitfully combine Wikipedia based measure, WordNet based semantic similarity and double checking model on the top extracted snippets of the queries in order to infer enhanced similarity measure. Future work involves study of algebraical and asymptotical properties of the elaborated measure as well as testing on alternative corpus. Especially, it is easy to see that expression (8) will require further refinements in the case where the presence of false negative is dominant in the dataset.

5 CONCLUSION

This paper contributes to the ongoing research of developing efficient tools for paraphrase detection. The approach advocates a web-based approach where the snippets of the search are analyzed using WordNet semantic based measure and Normalized-based distance Wikipedia based measure. The proposal has been designed in order to accommodate a prudent attitude like reasoning. The test using Microsoft Research Paraphrase Corpus has shown good results with respect to some of state of the art approaches. Although, the complexity of web search outcome is well documented, the proposal opens news ways to explore the timely availability of the search results by exploring the similarity of the search outcomes regardless of the accuracy of single search results.

REFERENCES

D. Bollegala, Y. Matsuo, and M. Ishizuka, 2007. "Measuring semantic similarity between words using web search engines," in Proc. of WWW '07, pp. 757–766

H. Chen, M. Lin, and Y. Wei, 2006. "Novel association measures using web search with double checking," in Proc. of the COLING/ACL 2006, pp. 1009–1016.

R. Cilibrasi and P. Vitanyi, 2007 "The google similarity distance," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 370–383.

R. Collobert and J. Weston, 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In ICML.

B. Dolan and C. Brockett, 2005. Automatically constructing a corpus of sentential paraphrases. In The 3rd International Workshop on Paraphrasing (IWP2005).

B. Dolan, C. Quirk, and C. Brockett 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, p 350, Morristown, NJ, USA. Association for Computational Linguistics.

C. Fellbaum, 1998. WordNet – An Electronic Lexical Database, MIT Press.

A. Fernando and M. Stevenson, 2008. A semantic similarity approach to paraphrase detection, Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics.

A. Islam, and Inkpen, D., 2007. Semantic similarity of short texts Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria, pp. 291-297.

R. Mihalcea, R., Corley, C., and Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity Proceedings of the National Conference on Artificial Intelligence (AAAI 2006), Boston, Massachusetts, pp. 775-780.

M. Sahami and T. Heilman, 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proc. of 15th International World Wide Web Conference*.

Z. Wu and M. Palmer, 1994. Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pages 133 –138, New Mexico State University,Las Cruces, New Mexico.

Y. Zhang and J. Patrick, 2005. Paraphrase identification by text canonicalization. In Proceedings of the Australasian Language Technology Workshop 2005, pages 160–166 Sydney, Australia, December.