International Journal of Hybrid Information Technology
Vol.8, No.5 (2015), pp. 31-40
http://dx.doi.org/10.14257/ijhit.2015.8.5.04

Semantic Similarity Computation Based on Multi-Features Fusion

Bo Ma, Yating Yang, Fan Zhao, Rui Dong and Xi Zhou

Research Center for Multilingual Information Technology, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences
mabo@ms.xjb.ac.cn

Abstract

Semantic similarity is a core technique of many NLP research fields. However, state-of-the-art metrics for semantic similarity computation often operate at different levels, e.g., words or sentences. In this paper, semantic similarity computation metrics are firstly introduced and the quality is measured in order to determine their advantages and limitations; then a new semantic similarity metric based on multi-features fusion is proposed. Distributed representations of words are used for alignment-based disambiguation operation, and Wikipedia tags are used to enhance the performance of our approach. The proposed metric is unsupervised, and can be applied at different levels e.g., single words or entire documents. The metric is evaluated on both English and Chinese datasets, it is shown that the precision and recall scores are higher than metrics which simply using knowledge base or distributed representation of words.

Keywords: *similarity; feature fusion; distributed representation; disambiguation; natural language processing*

1. Introduction

Semantic similarity is an essential technique for many applications in Natural Language Processing such as query expansion [1], Word Sense Disambiguation, and Question Answering. Evaluating semantic similarity is a central issue in many research areas such as Psychology, Linguistics, Cognitive Science, and Artificial Intelligence. Semantic similarity can also be exploited to improve accuracy of current Information Retrieval techniques [2].

Generally, supervised approaches and unsupervised approaches are two main categories to measure semantic similarity. The former consult human-built knowledge bases such as ontology. Compare with supervised approaches, unsupervised methods assume that the semantic similarity between words or texts can be extracted from the context by statistical analysis.

One problem faced by this line of work is that, by their nature, metrics for semantic similarity computation often have effect at different levels: methods for words, sentences and documents, which often make them inapplicable at the word or sentence level.

In this paper, we propose a unified approach for semantic similarity computation across multiple representation levels from words to documents, which offer one major advantage. The approach is useful independently of the input levels, which enables semantic similarity comparisons between different scales of texts.

The remainder of this paper is organized as follows: Section 2 provides some background information in the area of semantic similarity computation. In Section 3, the details of the proposed semantic similarity computation algorithm are discussed. Section 4 uses two datasets to compare the proposed method with several similarity metrics. Finally, we conclude the paper in Section 5; possible extensions of the proposed metric are also mentioned.

ISSN: 1738-9968 IJHIT

2. Related Work

Metrics that measure semantic similarity can be classified into two main branches: supervised metrics and unsupervised metrics.

2.1. Supervised Metrics

In our case, supervised metrics mean that knowledge bases are used as resources for computing the semantic similarity, and the most used resources are Word Net, HowNet and ontology.

Path-based metrics are the most popular methods which referring to the ontology. The value of semantic similarity is computed by considering the number of links between concepts. Obviously, the less number of links separating the concepts the more similar they are. This equation is modeled as follows [3]:

$$sim\left(c_i, c_j\right) = \begin{cases} e^{-\alpha l} \dfrac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, & c_i \neq c_j \\ 1 & c_i = c_j \end{cases} \tag{1}$$

Where l is the length of the shortest path between c_i and c_j, h is the level in the tree of the MSCA (Most Specific Common Abstraction) from c_i to c_j. The parameters α and β represent the contribution of the shortest path l and h. The optimal values for these parameters, determined experimentally, are: α=0.2 and β=0.6.

Information theoretic approaches are another branch to compute semantic similarity using ontologies, which employ the notion of Information Content (IC), which can be considered measuring the amount of information a concept expresses.

Resnik [4] was the first to use this method, the similarity depends on the amount of information two concepts have in common. The idea is modeled as follows:

$$sim_{res}\left(c_i, c_j\right) = \max_{c \in S\left(c_i, c_j\right)} IC\left(c\right) \tag{2}$$

Where $S\ (c_i,c_j)$ is the set of concepts that subsume c_i and c_j.

Pilehvar M T propose a unified approach for measuring semantic similarity called ADW, which uses Word Net as knowledge base and operates at multiple levels, all the way from comparing word senses to comparing text documents.

2.2. Unsupervised Metrics

Unsupervised metrics can be divided into three main categories: co-occurrence approaches, web based approaches and distributed representation. Co-occurrence approaches assume that the semantic similarity between words can be expressed by association ratio, which is a function of their co-occurrence. Web-based approaches use search engines to construct text corpus by exploiting the web source. And distributed representation is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships.

Mutual information (*MI*) measures the mutual dependence between the occurrence of words w_i and w_j [5], the maximum likelihood estimate of MI is:

$$I(X, Y) = \log \frac{\dfrac{\left|D \mid w_i, w_j\right|}{\left|D\right|}}{\dfrac{\left|D \mid w_i\right|}{\left|D\right|} \cdot \dfrac{\left|D \mid w_j\right|}{\left|D\right|}} \tag{3}$$

Cilibrasi and Vitanyi proposed a search engine based similarity method, called the Normalized Google Distance [6], to calculate the relationship between two words, defined as follows:

$$NGD\left(w_i, w_j\right) = \frac{\max\left\{\log f\left(w_i\right), \log f\left(w_j\right)\right\} - \log f\left(w_i, w_j\right)}{\log N - \min\left\{\log f\left(w_i\right), \log f\left(w_j\right)\right\}} \quad (4)$$

The attributes $f(w_i)$ and $f(w_j)$ represent the number of search results of the words w_i and w_j, respectively. The attribute $log\ f(w_i, w_j)$ represents the number of Web pages containing both w_i and w_j.

The word representation is computed by neural networks, the learned vectors explicitly encode many linguistic regularities and patterns. Many of these patterns can be represented as linear translations. This metrics provide state-of-the-art performance for measuring syntactic and semantic word similarities in several test sets [7].

3. Multi-Features Based Similarity Metrics

In this section, a similarity metric based on multi-features is proposed, it combines the advantages of both supervised approaches and unsupervised approaches, Sogou corpus and Wikipedia corpus are used as our training dataset for computing continuous vector representations of words. First TFIDF method is used to extract frequent words from texts to be compared, then the vector representations of frequent words is computed, word representation is also used to extend texts in the case of one text is much shorter than the other, the basic assumption behind this is that similarity of context represents similarity of meaning. Finally, Wikipedia tags are used as domain knowledge, and an unsupervised Wikipedia tags learning algorithm is proposed to improve the results of semantic similarity computing. Coefficients are assigned to each feature, and a score between [0, 1] is computed for each pair of compared texts.

3.1. Word Representations in Vector Space

The objective of training the model is to find vector representations that are useful for predicting the surrounding words in a sentence or a document. More formally, given a sequence of training words w1 w2, w3, ⋯ , wT, the goal is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\ne 0}\log p\left(w_{t+j} \mid w_t\right) \quad (5)$$

Where c is the size of the training context, the basic formulation defines $p\left(w_{t+j} \mid w_t\right)$ using the soft ax function:

$$p\left(w_O \mid w_I\right) = \frac{\exp\left(v_{w_O}^{'}\mathrm{T} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left(v_w^{'}\mathrm{T} v_{w_I}\right)} \quad (6)$$

Where v_w and $v_w^{'}$ are the “input” and “output” vector representations of word w, and W is the number of words in the corpus.

A computationally efficient approximation of the full softmax is the hierarchical softmax [8]. The main superiority is that it only needs to evaluate about $log_2\ (W)$ nodes instead of evaluating W output nodes in the neural network to obtain the probability distribution. The hierarchical softmax uses a binary tree to represent the output layer with the W words as its leaves; each word w can be reached by an appropriate path from the root of the tree:

$$p\left(w \mid w_I\right) = \prod_{j=1}^{L(w)-1} \sigma\left(\left\| n\left(w, j+1\right) = ch\left(n\left(w, j\right)\right)\right\| \cdot v'_{n(w,j)}\mathrm{T} v_{w_I}\right) \tag{7}$$

Where $\sigma(x) = 1/(1 + \exp(-x))$.

In very large corpus, we used a simple sub sampling approach to counter the imbalance between the frequent words and rare words: the probability of each word wi in the training set is computed by the equation:

$$P\left(w_i\right) = 1 - \sqrt{\frac{t}{f\left(w_i\right)}} \tag{8}$$

Where $f\ (w_i)$ is the frequency of word w_i and t is a threshold, typically around 10^{-5}.

3.2. Alignment-Based Disambiguation

Generally, semantic comparisons do not have senses of their lexical content annotated, however, traditional forms of word sense disambiguation are hard for short texts or single words because there is not enough contextual information presented to perform the disambiguation task. Therefore, we propose a novel alignment-based disambiguation algorithm that leverages the content of the paired item (words) in order to disambiguate each element.

Given two randomly ordered texts, we use word vector representation to seek the semantic alignment that maximizes the similarity of the context words in both texts. To find this maximum we use an alignment procedure that, for each word w_i in text T_1, assigns w_i to the word w_j that has the maximal similarity in the compared text T_2. Algorithm 1 formalizes the alignment process, which produces a sense disambiguated representation as a result.

Algorithm 1 Alignment-based Sense Disambiguation

```
Input: $T_1$ and $T_2$, the sets of words being compared
Output: $M$, the match word pairs for $T_1$ and $T_2$
M ← $\phi$
for each word $w_i \in T_1$
    $max\_sim \leftarrow 0$
    for each word $w_j \in T_2$
        $sim \leftarrow S(w_i, w_j)$
        if $sim > max\_sim$ then
            $max\_sim = sim$
    $M \leftarrow M \cup \{w_i, w_j\}$
return M
```

We use the two example texts T_1 and T_2 to illustrate the alignment-based disambiguation procedure, Figure 1 illustrates example alignments of the two texts.

T_1: 领导训斥了职员。

T_2 : 员工被老板批评了。

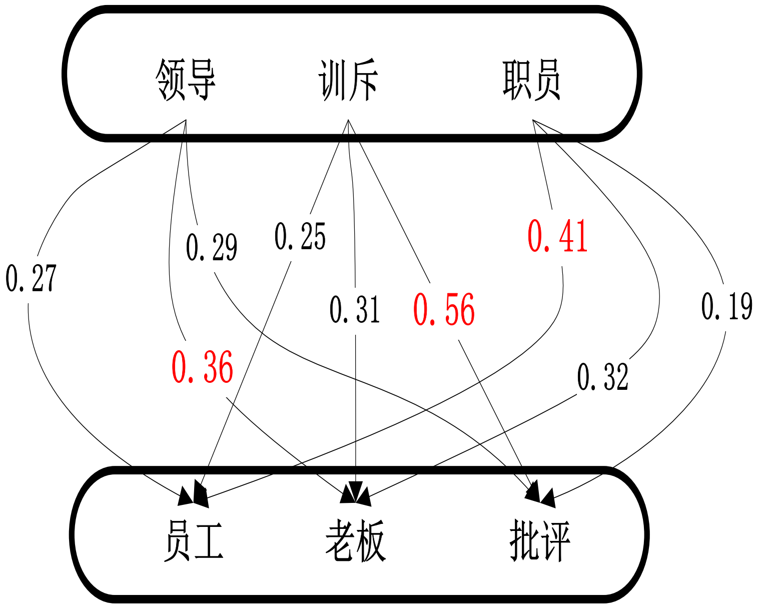

Figure 1. Example Alignments of Two Texts

Figure 1 shows the maximally-similar sense alignment of the words in T_1 and T_2. The algorithm generates the following alignment sets of matches:

$$M = \{\text{领导} \Leftrightarrow \text{老板，训斥} \Leftrightarrow \text{批评，职员} \Leftrightarrow \text{员工}\}$$

3.3. Unsupervised Wikipedia Tags Learning Algorithm

Wikipedia is a collaboratively edited, multilingual, free Internet encyclopedia that composed by tags or concepts [9]. Each tag is explained by an article, which corresponds to a set of categories. In this paper, we propose an unsupervised cluster-based algorithm which assigns Wikipedia tags to texts automatically.

We use v_{wi} to represent the distributed vector of word w_i, and Equation (9) to represent the distributed vector of a sentence:

$$v_S = \sum_{i=1}^{n} v_{w_i} / n \qquad (9)$$

Where S is a sentence and n is number of words of S.

```
Algorithm 2 Automatic Learning Algorithm
Input: $T = \{t_1, t_2, \cdots, t_n\}$, the sets of Wikipedia tags
       S, a sentence to be assigned tags
Output: $T_t = \{t_1, t_2, \cdots, t_m\}$, the match tags for S
1.  $T_t \leftarrow \phi$
2.  $Sim \leftarrow \phi$
3.  compute the distributed vector of S
4.  for each tag $t_i \in T$
5.      $sim_{S,t_i} = D(S, t_i)$
6.      add $sim_{S,t_i}$ to Sim
7.   $T_t \leftarrow$ Choose top 5from Sim
8.  return $T_t$
```

Then the semantic similarity computation metrics based on multi-features Fusion is proposed, we take all the features into account, using Sim_{BOW} to represent the similarity based on BOW (bag of words), using Sim_{w2v} to represent the similarity based on word representation, and Sim_{wiki} to represent the similarity based on Wikipedia tags.

$$Sim\left(t_i, t_j\right) = \alpha \cdot Sim_{BOW}\left(t_i, t_j\right) + \beta \cdot Sim_{w2v}\left(t_i, t_j\right) + \chi \cdot Sim_{wiki}\left(t_i, t_j\right) \tag{10}$$

α, β, χ are the coefficients and $\alpha + \beta + \chi = 1$. Determined experimentally, are: α =0.2, β =0.4, χ =0.4.

Then we use the Min-max normalization method to map the similarity value into [0, 1].

$$Sim\left(t_i, t_j\right) = \frac{Sim\left(t_i, t_j\right) - Sim\left(t_i, t_j\right)_{\min}}{Sim\left(t_i, t_j\right)_{\max} - Sim\left(t_i, t_j\right)_{\min}} \tag{11}$$

4. Experiments

4.1. Experiment Preparation

Data. Measuring semantic similarity of textual items has applications in a wide variety of NLP tasks. Microsoft Research Paraphrase Corpus (MSRP) containing 5800 pairs of English sentences which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship [10]. To measure the approach we proposed, we also use a Chinese Teaching Dataset (CTD) which contains 10000 pairs of mappings between teaching texts and resources. Table 1 and Table 2 show the statistics of the two datasets.

Table 1. Statistics of the MSRP Dataset

	MSRP	Training	Test
Total	5801	4076	1725
Semantic equivalence	3900	2753	1147
Non-semantic equivalence	1901	1323	578

Table 2. Statistics of the Chinese Dataset

	CTD	Training	Test
Total	10000	7500	2500
Semantic equivalence	8000	6000	2000
Non-semantic equivalence	2000	1500	500

Comparison Metrics. We compare our metric (BDV) against ADW [11] and word2vec [12]. ADW uses Word Net as knowledge base and operates at multiple levels, all the way from comparing word senses to comparing text documents, which achieves better results than top 3 systems of SemEval-2012. Word2vec is an efficient unsupervised implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words developed by Google.

4.2. Evaluation Metric

We choose precision, and recall, as the evaluation metrics.

$$\Pr ecision = TP / (TP + FP) \tag{12}$$

$$\operatorname{Re} call = TP / (TP + FN) \tag{13}$$

Where *TP* is the number of returned equivalent pairs of texts, *FP* is the number of returned pairs of equivalent texts which are not equivalent; and *FN* is the number of returned pairs of in equivalent pairs of texts that are actually equivalent.

4.3. Experimental Results

α is the threshold which changes from 0.0 to 1.0. The precision and recall of the three metrics on the MSRP are shown in Figure 2 and Figure 3. In Figure 2, the precision of BDV is a little higher than ADW, and they are both much higher than word2vec. That is because both BDV and ADW use some knowledge bases (Wikipedia tags and Word Net) to enhance the results, and BDV also uses word representation to enhance the performance. In Figure 3, BDV and ADW have better recall than word2evc, which is also because, the use of knowledge bases. The recall drops very fast when $\alpha \geq 0.6$, which suggests we can use $\alpha = 0.6$ as the threshold.

The precision and recall of the three metrics on the CTD are shown in Figure 4 and Figure 5. In Figure 4, the highest precision is achieved by BDV, and the lowest is ADW. That is because CTD is a Chinese dataset, and the knowledge base ADW used (Word Net) has no effect on it. Word2vec also has better performance than ADW, which is because word2vec is language irrelevant and the performance is relatively stable. In Figure 5, BDV and word2evc have better recall than ADW, which is also because the use of knowledge base (BDV) and the characteristics of language irrelevant (word2vec).

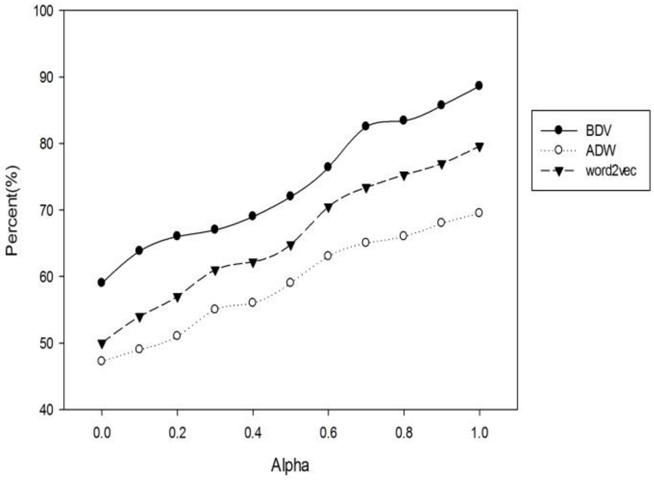

Figure 1. Precision with Different Threshold(Alpha) Based on CTD

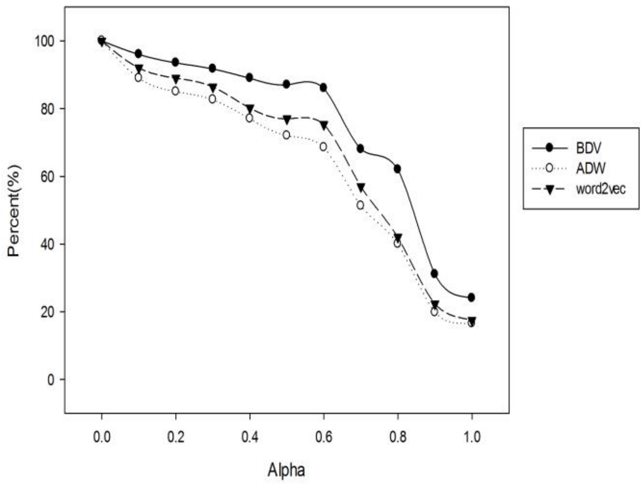

Figure 2. Recall with Different Threshold(Alpha) Based on CTD5.

Conclusion

In this paper, a new semantic similarity computation metric based on multi-features Fusion is proposed. The approach can be applied to both words and entire documents. The performance of the metric was evaluated on English sentence corpus MSRP and a Chinese dataset.

Good precision and recall scores were achieved by using the proposed metric. In the future, we plan to apply this metric into more NLP tasks such as recommendation systems and multilingual search engine.

Acknowledgements

Our thanks to Xi Zhou, Lei Wang, and all the members of the research center of multilingual information technology. This work is founded by West Light Foundation of The Chinese Academy of Sciences (No. XBBS201313), the National High Technology Research and Development Program of China (No. 2013AA01A607) and Xinjiang High-Tech R&D Program (No. 201230118-2).

References

[1] Y. Zhang, X. Wang, X. Wang, S. Fan, "Expanding User Intention by Type Similarity of Complex Questions", Journal of Computational Information Systems, vol. 5, no 3, **(2009)**, pp. 1245-1251.

[2] B. Ma, X. Zhou, Y. Yang, J. Zhou, "Uyghur Semantic Similarity Computation Based on Contextual Information in Web Documents", Journal of Computational Information Systems, vol. 8, no 2, **(2012)**, pp. 563-570.

[3] L. Yuhua, B. Za, D. Mclean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Trans on Knowledge and Data Engineering, vol. 15, no 4, **(2003)**, pp. 871-882.

[4] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", arXiv preprint cmp-lg/9511007, **(1995)**.

[5] K. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography", Computational Linguistics, vol. 16, no. 1, **(1990)**, pp. 22-29.

[6] P. Vitanyi, "Universal similarity", Information Theory Workshop, **(2005)** August 29-September 1.

[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space", arXiv preprint arXiv: 1301.3781, **(2013)**.

[8] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model", Proceedings of the international workshop on artificial intelligence and statistics, **(2005)**, pp. 246–252.

[9] X. Hu and X. D. Zhang, "Exploiting Wikipedia as External Knowledge for Document Clustering", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2009)**, pp. 389-396.

[10] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploring massively parallel news sources", Proceedings of the 20th International Conference on Computational Linguistics, **(2004)**.

[11] M. T. Pilehvar, D. Jurgens, R. Navigli, "Align, disambiguate and walk: A unified approach for measuring semantic similarity", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **(2013)**, pp. 1341-1351.

[12] T. Mikolov, I. Sutskever, K. Chen, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems, **(2013)**, pp. 3111-3119.

Authors

Bo Ma, receives his Ph.D. from Graduate University of Chinese Academy of Sciences and is on the way of his research work. He is currently an Assistant Professor at Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China.

Email: mabo@ms.xjb.ac.cn

Yating Yang, receives her Ph.D. from Graduate University of Chinese Academy of Sciences and her research interest is machine translation. She is currently an Associate Professor at Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China.

Email: yangyt@ms.xjb.ac.cn

Fan Zhao, receives his master's degree from Graduate University of Chinese Academy of Sciences and he is a Ph.D. candidate. He is currently an Assistant Professor at Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China.

Email: yanyushu@gmail.com

Rui Dong, receives his master's degree from Graduate University of Chinese Academy of Sciences and he is a Ph.D. candidate. He is currently an Assistant Professor at Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China.

Email: dongrui@ms.xjb.ac.cn

Xi Zhou, receives his Ph.D. from Graduate University of Chinese Academy of Sciences and his research interest is multilingual information processing. He is currently a Professor at Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China.

Email: 0716genius@163.com