Combination Features for Semantic Similarity Measure

Dat Huynh, Dat Tran, Wanli Ma

Abstract—**Computing the semantic similarity between words is one of the key tasks in many language-based applications. Recent work has focused on using contextual clues for semantic similarity computation. In this paper, we propose a method to the measure semantic similarity between words using plain text contents. It takes into account information attributes (local) and topic information (global) of words to disclose their semantic similarity scores. The method models the representation of a word as a high dimensional vector of word attributes and latent topics. Thus, the semantic similarity between two words is measured by the semantic distance between their respective vectors. We have tested the proposed method on WordSimilarity-353 dataset. The empirical results have shown the combination features contribute to improve the semantic similarity results the dataset in comparison with previous work on the same task using plain text contents.**

Index Terms—**Semantic Text Analysis, Semantic Similarity, Semantic Relatedness, Distributed Representation.**

I. Introduction

In many language-based applications, such as semantic search, word disambiguation, and text clustering, computing the semantic similarity between words is the crucial and fundamental task. Previous work in the field is categorized as the knowledge-based and content-based approaches. While the knowledge-based approaches utilise the embedded knowledge in corpora such as Wordnet, Wikipedia links, Flickr image tags, and Del.icio.us bookmarks, the content-based methods rely on the large amount of plain text contents existed to measure the semantic similarity.

Vector Space Models (VSMs) remain the favourite model for word meaning representation. In knowledge-based approaches, Explicit Semantic Analysis (ESA) was proposed to construct meanings of a word as a high dimensional vector of Wikipedia concepts [1]. Silent Semantic Analysis (SSA) [2] utilizes Wikipedia Concepts in contexts to model word meanings using VSMs.

However, the contribution of VSMs are differently in the content-based approaches. VSMs were used to model the representation of a word as a high dimensional vector of its context window patterns [3]. Word meanings in were also represented in a high dimensional vector of latent concepts using word-document associations [4], [5]. Relational patterns between word pairs were used as vector features for measuring semantic similarity between pairs of words [6]. Lexical syntactic patterns of words in contexts over a parsed corpus were used to measure the words' similarity [7].

It has been recognized that the semantic meanings of a word are determined by its surrounding contexts, and the similarity between words could be determined by the common of their respective contexts. In this paper, the meanings of a word are examined by its attributes. Given two focus words, the similarity between them is determined by the common of their attributes. Moreover, a word itself also has multiple facets belonging to different topics. Thus, we examine if the common topics (in the global sense) of individual words would contributes to measure semantic similarity of words. Finally, the combination about the local features in contexts (attributes) and global features (topics) would be also considered in word semantic similarity.

Different from other previous approaches, the contribution of this work is to examine the effectiveness of local attributional features and global topic features in word representation. This also is tested by the task of semantic similarity and compared with other related work on the same task using only plain texts from a large text repository. Our experiment has confirmed the promising results on those kinds of features in comparison with other previous work on the same tasks.

II. Word Representation Using Attributional Features

It has been confirmed that meanings of a word is determined by its surrounding contexts. The surrounding contexts include syntagmatic relations and paradigmatic relations, which jointly describe the meanings in different aspects [8]. While paradigmatic relations hold the meanings over long distant relations in local contexts, the syntagmatic relations contain the meanings when the word interacts with its adjacent neighbours. For instance, given the local contexts as in the sentences: "*the bowl of Pho is served in a restaurant*"; "*Noodle soup is referred in Asian restaurants*". The meanings of the unknown word "*Pho*" could be inferred by its syntagmatic relations with surrounding words such as "*bowl, restaurant, serve*", while paradigmatic relations express meanings of "*Pho*" via words that on the same categories such as ,"*noodle*", "*soup*".

In this work, we mainly focus on exploring the contribution of the syntagmatic relations in describing word meanings. Given a word in contexts, we considered relational words in syntagmatic relations as attributes of that word. Moreover, a word appeared in different contexts returns its attributes with different level of importance contributed to its meanings. Thus, the combination of the attributes would be used to disclosed the meaning of the word in multiple contexts. Here, we used a high dimensional vector of attributes to model meanings of the word. Given a focus word w_i and its VSM representation $v(w_i)$ is described as follows:

Dat Huynh is a PhD candidate at the University of Canberra, Australia. Email: Dat.Huynh@canberra.edu.au. Dat Tran is an Associate Professor at the University of Canberra, Australia. Email: Dat.Tran@canberra.edu.au. Wanli Ma is an Assistant Professor and an Academic Program Leader at the University of Canberra, Australia. Email: Wanli.Ma@canberra.edu.au

$$v(w_i) = \langle w_i^1, w_i^2, \ldots, w_i^n \rangle \tag{1}$$

where w_i^k, the level of importance reflecting the semantic association between the word w_i and its attribute w_k, and n is the number of distinct words in the given text repository. To measure the level of importance of each attribute of the word w_i, we used point-wise mutual information (PMI) [9], which computes the degree of information value (association) between two different words in their appearance contexts. The information value w_i^k of the pair of words (w_i, w_k) is measured as follows:

$$w_i^k = \log \frac{p(w_i, w_k)}{p(w_i)p(w_k)} \tag{2}$$

$$p(w_i, w_k) = \frac{d(w_i, w_k)}{\sum\limits_{i,k=1\ldots n} d(w_i, w_k)} \tag{3}$$

$$p(w_i) = \frac{\sum\limits_{k=1\ldots n} d(w_i, w_k)}{\sum\limits_{i,k=1\ldots n} d(w_i, w_k)} \tag{4}$$

where $d(w_i, w_k)$ is the number of times that w_i and w_k co-occur in syntagmatic relations.

III. Word Representation Using Latent Topic Features

ESA and other similar methods have successfully used explicit topics such as Wikipedia concepts to describe word meanings. The method relies on the special structure of Wikipedia links, which hardily adapts to different domains as well as languages. In this work, we expect to use the latent topics instead, which could be inferred from a generative topic model operated on a large text corpus. Several variants of topic model has been proposed such as Latent Semantic Analysis (LSA) [4], Latent Dirichlet Allocation (LDA) [10]. They are all based on the same fundamental idea that documents are mixtures of topics where a topic is a probability distribution over words, and the content of a topic is expressed by the probabilities of the words within that topic. In our experiment, we used LDA as the background topic model in building features for word representation. LDA performs the latent semantic analysis to find the latent structure of "topics" or "concepts" in a text corpus.

Given a focus word w_i and a latent topic t_j, the topic model returns the probability m_i^j that w_i belongs to the particular topic t_j. Thus, the topic representation of the word w_i is considered as a vector of latent topics, where each value of the vector is represented for the probability that w_i belongs to particular topics t_j $(j = 1 \ldots k)$.

The topic representation of the word w_i is described as follows:

$$u(w_i) = \langle m_i^1, m_i^2, \ldots, m_i^k \rangle \tag{5}$$

where k is the number of latent topics. The vector $u(w_i)$ is used to describe the meanings of the word w_i using latent topic information.

IV. Representation Using Word Topic Combination Features

Given w_i as a focus word, meanings of the word w_i are represented as a n dimensional vector $v(w_i)$ of attributional words denoted $w_1 \ldots w_n$ (see formula 1). Meanwhile, the focus word w_i is also represented as a vector $u(w_i)$ of latent topics denoted $t_1 \ldots t_k$ (see formula 5). We suggest a straightforward way to represent the word meanings using both attributional information and latent topic information. The composition vector representation $c(w_i)$ of the word w_i is the linear concatenation of the attributional feature vector $v(w_i)$ and the latent topic feature vector $u(w_i)$ as:

$$c(w_i) = \langle \alpha w_i^1, \ldots, \alpha w_i^n, \beta m_i^1, \ldots, \beta m_i^n \rangle \tag{6}$$

where n is the number of word features and k is the number of latent topics. The α and β are parameters reflexing the contribution of each kind of features over the tasks of similarity measure.

V. Word Semantic Similarity

The proposed content-based method of semantic similarity was constructed using two different group of features: words in contexts as attributes and latent topics. These groups of features were tested separately and collectively. Therefore, the following pre-processing steps were undertaken:

1) *Word Attribute Extraction*: Attributional features of each words need to be extracted from a plain text repository. To keep the approach as simple as possible and to maintain the ability to adapt to different languages, we attempt to use N-gram technique to extract pairs of co-location words in a certain window size (W) local contexts. For each pair, one word is considered as the focus word, while the other is considered as its attributes. Although the N-gram technique is not great choice to syntagmatic relations, we believe that their redundancy information incorporated with suitable weighting filter parameters could help to leave out parts of non-essential information.
2) *Word Meaning Representation*: The representation of a word using attributional features is constructed after applying the first frequency filter (FF) on each pair and the second information value filter (IVF) on each pairs. The remained pairs after filtering out will be represented in VSMs for word meaning representation.
3) *Distance Measure*: To measure the semantic similarity between two words, we directly used the standard *Cosine* distance measure on the representation vectors. Given two words w_i and w_j, the semantic similarity between them is computed as:

$$sim(w_i, w_j) = \frac{v(w_i) \times v(w_j)}{\|v(w_i)\| \|v(w_j)\|} \tag{7}$$

VI. Implementation Details

A. Text Reponsitory

We used Wikipedia English XML dump of October 01, 2012. After parsing the XML dump[1], we obtained about

[1] We used Wikiprep as the main tool to convert Wikipedia format to XML plain text, http://sourceforge.net/projects/wikiprep/

13GB of text from $5,836,084$ articles. As we expect to have a reasonable large amount of text data to increase the coverage of the method, we used first $1,000,000$ articles for our experiments.

To build the representation for each word, we used N-Gram model to extract pairs of words within a windows size of $W = 3$ words from the Wikipedia plain texts after removing stop-words. Then, we applied the stemming technique [11] to all the extracted words. We finally obtained over $224M$ unique pairs overall. However, there is the large number of rare pairs with very low frequency. We applied the first frequency filter (FF=2) to remove non-essential word association in pairs. Additionally, we applied the second information value filter (IVF) on each pair. We expect to monitor the influence of IVF on the performance of the similarity measure (see Table II). Only pairs have their information values equal or above the IVF will be retained to form the representation of words.

To extract latent topic features, we used the first $100,000$ documents to build LDA training model. The reasons for us to choose this smaller amount of documents as LDA training phrase was time consuming with large amount of documents. we expected to reduce the number of input documents and kept the word dictionary was relatively large to cover most of the expected words. The plain text from these documents was removed stop-words and and stemmed before training. We obtained $190,133$ unique words from the given set of documents after pre-processing step. To build the LDA training model, we used GibbsLDA++ implementation [12] with its default configuration except $ntopic = 1,000$ as the number of expected latent topics.

Finally, as the different ranges of values on vector representation using attributional features and using latent topic features, the performance of the combination feature-based method would be affected by the $\frac{\alpha}{\beta}$ ratio. After experimented with the same method on an independent test, we selected $\frac{\alpha}{\beta} = 0.002$ for our experiments.

B. Dataset

WordSimilarity-353 [13] dataset has been one of the largest publicly available collections for semantic similarity tests. This dataset consists of 353 word pairs annotated by 13 human experts. Their judgement scores were scaled from 0 (unrelated) to 10 (very closely related or identical). The judgements collected for each word pair were averaged to produce a single similarity score. Several studies measured inter-judge correlations and found that human judgement correlations are consistently high $r = 0.88 - 0.95$ [14], [13]. Therefore, the outputs of computer-generated judgments on semantic similarity are expected to be as close as possible the human judgment correlations.

C. Evaluation

In this section, we present our experimental results and compare with other work of the same task[2].

On the standard WordSimilarity-353 dataset, Table I shows the correlation between computer-generated results and human judgements. Compare with other work on the same task using only plain texts, our attributional features and combination features outperformed most of the related method on the same category. Furthermore, the proposed method has achieved the second best results on the similarity measure using the same kind of data after the multi VSMs (multi-prototype) representation [15].

TABLE I
THE COMPARISON BETWEEN OUR RESULTS AND OTHER WORK IN CONTENT-BASED SEMANTIC SIMILARITY TASKS ON WORDSIMILARITY-353 DATASET USING SPEARMAN'S RANK CORRELATION (ρ).

Algorithm	ρ**×100**
Syntactic-based [7]	34.80
LSA-based [4]	58.10
Topic-based [5]	53.39
Multi-Prototype [15])	76
Single-Protoype [15]	53
Learned Features [16]	49.86
Context Window Patterns [3]	69
Word Features (IVF=1.0)	71.09
Topic Features	67.01
Combination Features (IVF=1.5)	**73.67**

Topic features on Wikipedia texts outperformed to those tested on other datasets [5] and also produced better performance to the long-standing LSA. This again confirmed the advantage of LDA on semantic similarity tasks as well as the important contribution of using larger and richer plain text repository for building topic model.

Different the complexity of building topic features, building attributional features is very straightforward support, but produces promising results compared to previous attempts on the similar features such as pure syntactic-based features [7], context window pattern features [3], as well as features that automatically learned from the nature patterns on texts [16]. One of the major differences from our extracted features is that the selection procedure heavily is relied on the operation of the information value filter (IVF) (see Figure 1) as while most of the other work tends to use pair frequency filter (FF). As our observation, FF has the minor effect in the performance of the task of similarity measure. Although attributional features and latent topic features have contributed differently in the tasks of similarity measure, there is a possibility to combine them together in to a better representation. The correlation result from the combined features ($\rho = 0.7367$) has confirmed the advantage of this kind of features on the standard WordSimilarity-353 dataset.

Parameter Turning: As a content-based method, the performance of our method is pretty much depending on the given plain text corpus. We expect to see how the performance could be effected by the data input adjustments. In our experiment, we kept stably the first filter ($FF = 2$) of pair frequency. We suspected that with a large amount of input data, the pairs with lower than that frequency would not be useful in general. Therefore, we adjusted the second filter, information value filter (IVF). The Figure 1 shows how the IVF does affect the performance of our methods on three different kinds of features. With different kinds of features and with different amount of data fed to the system, the correlation results have been changed depending on IVF values. However, the Figure 1 also confirmed that the combination features outperforms over the word features

[2] The experiment results can be found at http://137.92.33.34/IAENG2014Data

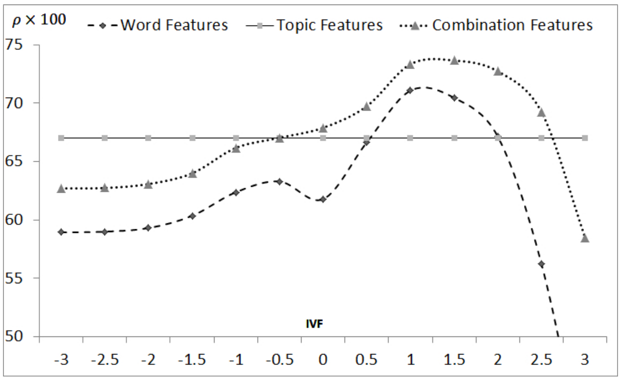

Fig. 1. This is the visualization of Table II. The combination features outperformed the word features in all tests with different information value filter (IVF). The training data for topic model does not involve IVF at all (ρ=0.6701)

on the task of semantic similarity regardless of the IVF values. As the training data fed to LDA topic model does not involve IVF, the results based on this kind of features is stable (ρ=67.01) during the tests. The Figure 1 also shows that when the large number of pairs has been removed (as IVF increased), the performance of the method significant dropped down. At the point of $IVF = 3$, there was only about $6M$ pairs fed to the system in comparison with about $27M$ and $18M$ pairs when IVF=1.0 and $IVF = 1.5$ respectively.

TABLE II
THE CORRELATION RESULTS WITH DIFFERENT INFORMATION VALUE FILTER (IVF) TESTED ON WORDSIMILARITY-353 BENCHMARK USING SPEARMAN'S RANK CORRELATION (ρ). THE COMBINATION-FEATURE-BASED METHOD OUTPERFORMED OVER THE METHOD BASED ON ATTRIBUTIONAL FEATURES REGARDLESS IVF.

IVF	Word features	Combination features
-3.0	58.95	62.72
-2.5	59.01	62.75
-2.0	59.32	63.09
-1.5	60.37	64.01
-1.0	62.39	66.19
-0.5	63.31	67.05
0.0	61.80	67.91
0.5	66.67	69.76
1.0	**71.09**	73.36
1.5	70.47	**73.67**
2.0	67.14	72.74
2.5	56.23	69.25
3.0	38.78	48.48

High Dimensional Vector: Although the proposed method presented the promising results compared with other methods, certain issues could be earning worth thoughts.

Firstly, the method has used full range of word and topic features. This yields very high dimensional vectors. For instance, in the case of $IVF = -3$ we had $485,513$ purely word features and $1,000$ additional latent topic features for the combination feature test. Similarly, in the best situation where $IVF = 1.5$, there was not significant change of the number of features. It was about $468,617 + 1,000$ word and topic features. We have tried to reduce the number of features by select top highest values of word and topic features separately. However, any attempts to select the top highest values on the vectors dealing with decreasing of similarity correlation score overall.

Secondly, the method has shown the very simple way of combining different sets of features. We were using the linear combination of word and topic features depending on α and β parameters ($\frac{\alpha}{\beta} = 0.002$ was used in the tests). We believe that different techniques could be attempted to find the best possible way to combine these kinds of features.

VII. CONCLUSION

We have presented an approach for semantic similarity measure. The method takes into account the word attributes in local contexts and latent topics information from global contexts. The experimental results have shown the positive contribution of the attributional features and topic features in comparison with those previously tested. Especially, the combination between attributional and topic features on word representation yields the outperformance results to most of the content-based methods on WordSimilarity-353 dataset.

REFERENCES

[1] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJCAI*, vol. 7, 2007, pp. 1606–1611.

[2] S. Hassan and R. Mihalcea, "Semantic relatedness using salient semantic analysis." in *AAAI*, 2011.

[3] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 19–27.

[4] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.

[5] G. Dinu and M. Lapata, "Measuring distributional similarity in context," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1162–1172.

[6] P. D. Turney, "Mining the web for synonyms: Pmi-ir versus lsa on toefl," in *Proceedings of the 12th European Conference on Machine Learning*, 2001, pp. 491–502.

[7] D. Lin, "An information-theoretic definition of similarity." in *ICML*, vol. 98, 1998, pp. 296–304.

[8] M. Sahlgren, "The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces," Ph.D. dissertation, Stockholm, 2006.

[9] I. Dagan, S. Marcus, and S. Markovitch, "Contextual word similarity and estimation from sparse data," in *Proceedings of Association for Computational Linguistics*. Association for Computational Linguistics, 1993, pp. 164–171.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] C. J. Van Rijsbergen, S. E. Robertson, and M. F. Porter, *New models in probabilistic information retrieval*. Computer Laboratory, University of Cambridge, 1980.

[12] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 91–100.

[13] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.

[14] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[15] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 873–882.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.