Measuring Semantic Textual Similarity using modified Information Content of WordNet and Trigram Language Model

Goutam Majumder, Partha Pakray, David Eduardo Pinto Avendano
National Institute of Technology, Mizoram
India
goutam.nita@gmail.com, parthapakray@gmail.com, davideduardopinto@gmail.com

ABSTRACT: *The proposed method is developed for measuring the textual similarity using WordNet taxonomy and information content. It uses the taxonomy knowledge and merge this information with an n–gram based language model (n = 3). The proposed method considers WordNet synsets for lexical relationships between the words and language model for information content value between the two words of the different class. Finally, a similarity score is generated between two sentences by measuring maximum weight shortest path of the graph. To evaluate the system SemEval 2015 training dataset is considered and the high correlation coefficient of 0.885 has been achieved.*

Keywords: Taxonomy, WordNet, Trigram, Information Content, Named Entity, Semantic Similarity, Natural Language Processing

Received: 14 May 2017, Revised 12 July 2017, Accepted 19 July 2017

1. Introduction

In Natural Language Processing (NLP) semantic similarity plays an important role and one of the fundamental tasks for many NLP applications and its related areas. During the evolution of Semantic Textual Similarity (STS), it is defined by a metric over a set of documents where the idea is to find the similarity between them [1], [2]. Similarity between the documents is measured based on the direct and indirect relationships among them. These relationships can be measured and recognized by the presence of semantic relationships. Identification of STS in short text was proposed in 2006 [3], [4]. After that focus of the similarity measure was shifted on large documents (or individual words). Since its inception, the problem has been seen a large number of solutions in a relatively small amount of time. The central idea behind most of the solution was the identification and alignment of semantically similar or related words across the two sentences and the aggregation of these similarities to measure the overall similarity [3], [5], [6].

One of the major goals of STS task is to create a unified framework by combining several independent semantic components to find their impact over several NLP tasks. Developing such framework is an important research problem for many NLP applications such as Information Retrieval (IR) text summarization [7], [8], question answering [9], relevance feedback and text classification [10], word sense disambiguation [11], and extractive summarization [12].

This proposed method is developing for measuring the textual similarity between texts based on language model and WordNet

taxonomy. The rest of paper is organized as follows: Section II presents a brief overview of the related work. Proposed method is discussed in Section III. Experimental results and evaluation are discussed in Section IV and finally future work and conclusion has been drawn in Section V.

2. Related Work

Measuring semantic similarity between text snippets is categories into following groups: (i) topological approaches; (ii) statistical similarity approaches; (iii) semantic-based approaches; (iv) vector space approaches; (v) alignment approaches and; (vi) machine learning approaches. Among these methods, taxonomy based approaches play an important role to understand the intended meaning of an ambiguous word like 'bank'. A bank can be a financial or may be a river bank and difficult to process computationally. For many NLP related task, it is important to understand the semantic relation between words/ concepts. To decompose such systems, it is important to work with word level relations and those can be considered as hierarchical, associative and equivalence.

In many cases determining the intended meaning of an ambiguous word is difficult for human and it is quite difficult to process automatically also. This ambiguity can be estimated by considering the following relationships among the words or concepts: (i) hierarchical (e.g. IS–A or hypernym–hyponym, part-whole etc.), (ii) associative (e.g. cause–effect) and, (iii) equivalence [13]. Among these IS–A relation has been widely used and studied, which maps to the human cognitive view of classification (i.e. taxonomy). The IS–A relation among the concepts has been suggested and employed as a special case of semantic similarity of distance [14]. Semantic similarity can be estimated by defining a topological similarity by using the ontological relationships, which measures the distance between the terms and concepts.

Islam et al., [15] proposed an unsupervised approach for measuring the similarity of texts using the corpus-based n– gram word similarity. First, they identified the size of n–gram for better word similarity task. In literature, it was observed by Kaplan [16], that word sense of the either side of the word is better than the words those are preceding or the following. So they consider the size of n as three (3) over bi–grams and other language model [17].

Malandrakis et al., estimate the semantic similarity between two sentences using regression models. To measure the similarity three features have been considered: 1) for lexical matching n-gram hit ratio was taken; 2) lexical similarity for non–matching words; and 3) sentence length is also taken into account. Authors also considered information metric for computing the lexical semantic similarity via co–occurrence counts over a web corpus [18].

To find the similarity for phrases or sentences a random walk over a graph was proposed by Ramage et al. In this work local semantic information and semantic resources of WordNet have been combined together. The semantic signature generated by random walk was compared to another such distribution to get the overall similarity score [19].

Resnik, P proposed another such method for identification of semantic similarity in a taxonomy based on the notion of Information Content (IC) [20]. Similarity between two words/ concepts was evaluated by considering the common information between them. For this task, a set of fifty thousand (50,000) nodes from WordNet taxonomy of noun classes had been considered. To calculate the frequencies of concepts Brown Corpus of American English (having 1000, 000 words) [21] was considered.

3. Proposed Method

In this paper, a language model based semantic network has been proposed to find the semantic similarity between two text snippets. Unlike the work reported in [20] and [22] STS model, the proposed method does not stick to the similarity between the concepts those poses IS–A relationship using information content (IC) value. In this task, words those are not sharing any relationship we calculate IC value using tri–gram language model over a corpus.

3.1 WordNet based Textual Similarity

In the first step two sentences S_1 and S_2 have been analyzed to extract all the WordNet synsets related to them. For each WordNet synsets, noun synsets are considered and generates two sets of synsets C_1 and C_2 of sentence S_1 and S_2 respectively. For the synsets found other categories the IS– A taxonomy relationship is considered. Intuitively, one key to the similarity of two concepts is the extent to which they share information in common. In taxonomy direct relation between two concepts can be

found by an edge counting method. In this method, if the minimal path between two nodes is long, that means it is necessary to go high in the hierarchy to find a least upper bound. For instance in WordNet, two concepts like 'NICKEL' and 'DIME' both subsumes 'COIN' sub–class, whereas 'NICKEL' and 'CREDIT CARD' share a common super class 'MEDIUM OF EXCHANGE' [23].

Taken C_1 and C_2 as the set of concepts of sentences S_1 and S_2, with $|C_1| \geq |C_2|$, the conceptual similarity between S_1 and S_2 is measured as follows:

$$sim(S_1, S_2) = \frac{\sum_{c_1 \in C_1} \max_{c_2 \in C_2}}{|C_1|} \tag{1}$$

where c_1, c_2 represents the each concept of the synsets and $s(c_1, c_2)$ represents measuring of conceptual similarity. For this method the conceptual similarity is measured by considering the IC value between the concepts. For this task the probability has been associate with the taxonomy. The value of IC of a class *c* is obtained by estimating the probability in a large corpus with a function p : $C \rightarrow [0; 1]$, if $c \in C$, $p(c)$, is the probability of encountering an instance *c*. Considering the notation of information theory [24], IC of a class c can be calculated as follows:

$$IC(c) = \log^{-1} p(c) \tag{2}$$

Quantifying information content in this way: if the probability increases, its information content decreases. It means that if there is a unique top in the tree, then its probability is 1 so the information content is zero (0) and the similarity of two concepts can be calculated as follows:

$$sim(S_1, S_2) = \max_{c \in Set(C_1, C_2)} [-\log p(c)] \tag{3}$$

2.2 Information Content based Similarity

To implement the information content model [20], consider the WordNet (version 2.0) having fifty thousand (50000) nodes [13], where taxonomy of concepts represented by nouns and compound nominals [25]. Before implementing IC, we need to define two concept sets as *words(c)* and *classes(w)*. *Words(c)* is the set of words subsumed by the *class c* and *classes(w)* is defined as the classes in which the word *w* is contained. The class c can be seen as a sub–tree in the whole hierarchy and *classes(w)* is the set of possible senses that the word *w* has:

$$classes(w) = \{c | w \in words(c)\} \tag{4}$$

A simple class/ concept frequency formula defined by [20] as follows:

$$freq(c) = \sum_{w \in words(c)} freq(w) \tag{5}$$

and the similarity metric is as follows:

$$s(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2IC(\dot{c_1}, c_2)} \tag{6}$$

So the similarity between two sentences S_1 and S_2 is calculated as follows:

$$sim(S_1, S_2) = \frac{1}{2}\left[\frac{\sum_{w \in S_1} \max_{w_2 \in S_2} ws(w, w_2) * idf(w)}{\sum_{w \in S_1} idf(w)} + \frac{\sum_{w \in S_1} \max_{w_2 \in S_2} ws(w, w_2) * idf(w)}{\sum_{w \in S_1} idf(w)}\right] \tag{7}$$

where *idf(w)* is calculated as the inverse document frequency of word w, by considering the WordNet 3.0 frequency count. In this case one word have more sense, so the similarity can be determined by the best similarity value among all the class pairs which their various senses belong to [13]:

$$sim(w_1, w_2) = \max_{c_1 \in sen(w_1)\ c_2 \in sen(w_2)} [sim(c_1, c_2)] \quad (8)$$

where *sen(w)* denotes the set of possible senses for word *w*.

3.3 Named Entity Overlap

For this task we consider Stanford Named Entity Recognizer [26] and we consider seven (7) different Named Entity (NE) classes as: Location, Percent, Organization, Date, Person, Time and Money. After that per-class overlap measure has been calculated as follows:

$$OVLP_{ner}(S_1, S_2) = \frac{2 * |NER_{S_1} \cap NER_{S_2}|}{|NER_{S_1}| + |NERS_2|} \quad (9)$$

where NER_{S1} and NER_{S2} represents the set of Named Entities identified in S_1 and S_2 respectively.

3.4 Tri–gram and Taxonomy based Similarity

In this method main contribution is that if method is unable to find any IC value between two concepts of two sentences then tri–gram frequency count in corpus has taken into account as IC value. The proposed system can be brought down into following stages:

1) Stage 1: In any language processing it is important to remove all the stop words before start any semantic similarity task. Initially all the stop words, have stored in a Java array and after that all the words of S and T is considered one after another for identification. Although stop words are most commonly used words but there is no universal list available for all language processing task [1]. These identified stop words are ignored during similarity stage.

2) Stage 2: In first step, Peen Treebank tag set [27] has been used to label the words for POS information, which is most commonly used syntactic information. Further these tag sets along with words have been input to the system to generate the parse tree of the sentences.

3) Stage 3: To generate the parse tree top down parsing has been followed by considering its advantages over bottom up parsing. For parsing all phrase structure of grammar has been considered. After that identified phrase structures have been used to generate the top–down parse tree.

4) Stage 4: In this stage a multi–stage (equal to level of the tree) undirected weighted graph has been designed by considering the parse tree along with other statistical information found in the previous stage. Following characterises have been considered for the graph construction:

POS: All the stop words based on its POS information isn't consider, when two words are found same in two parse tree at same level.

Node Depth: Starting from the root node S all possible paths have been considered till the search ends with a word/ concept at higher level (i.e. leaf node) of the tree. The depth of any word is consider in the similarity measuring stage when a word has been found in both parse tree are same and having same POS information. String Matching: If any word has been found in the parse tree of S1 and S2, which possess 'NNP' as POS tag then string matching algorithm has been found to assign the weight value of the link.

[1] http://xpo6.com/list-of-english-stop-words/

5) Stage 5: After the completion of graph construction stage weight has been measured between the nodes of two graphs. Assigning of weight has been performed under the following criteria's:

If POS tag has been found different of two classes of the same level leaf node then WordNet taxonomy relationship has been considered. To calculate the information content i.e. weight wi of the link the negative logarithm of the conditional probability (see Equation 2) as well as the argument of information theory is considered.

If POS tag is different but strings are matched than two different weight values have been calculated.

$$w_i^1 = sim\,(c_1, c_2) \tag{10}$$

and

$$w_i^2 = freq_{counts}\left(\frac{c_i}{N}\right) \tag{11}$$

where c_1 and c_2 represents two concepts of two parse tree at the same level. *N* represents a total number of words along with POS tag from a large text corpus and c_i represents a total of class *c* and $freq_{counts}$ counts the number of occurrences of each class. The final is taken as the maximum of w_i^1 and w_i^2 .

• If no condition matched and the phrase has found as noun class and words are proper noun then no weight has been measured for the link between the current node and the proper noun node.

• Finally, similarity has been calculated as the minimum distance path while considering maximum weight policy. After that, an average has been calculated by summing of all weights of links starting from start node to till the leaf node.

4. Experimental Results

In order to evaluate our text similarity measure, a pair of 50 sentences has been taken from SemEval training 2015 dataset. For this task, we analyzed four different methods to check the efficacy. We achieved 0.46 as similarity score using WordNet Taxonomy and we consider it as method 1. For this task WordNet version, 2.0 is considered.

For another method, we improved the similarity score using information content. For this task highest score is 0.78. In this method, we change the way information value have been measured, as compare to the methods introduced in [13], [20], and [22]. In this task n–gram language model was imposed and in this case size of *n* is 3.

In the third method, similarity score was generated by considering Named Entity Overlapping method. In which we calculate a per–class overlap measure where same string having different name entity tag does not match. Although it increases the score of similarity over the method 1, it lowers the score than IC (method 2).

In the final method where we combined all three features from three methods with modified WordNet taxonomy based on tri–gram language model for calculating IC value between two concepts of S_1 and S_2. In this method, we achieve high similarity score 0.885, which outperforms other taxonomy based methods discussed in [13], [20], [22], [28].

5. Conclusion and Future Work

From this work, it is clearly understood that node based approach fully depends on information content value between two nodes while distance based approach depends on the depth of the semantic network. On the other side hybrid method works with weight value between child and parent node to find the similarity between two classes. The proposed method, which uses the uni–gram model and hybrid method for measuring the weight between two nodes while using the advantages of WordNet information like node-based and distance-based approach. Finally, the minimum path and maximum weight have been considered

for generating the similarity score between the two texts. In future, we have planned to compare the proposed method with the other method based on machine learning algorithm.

References

[1] Corley, C., Mihalcea, R. (2005). Measuring the semantic similarity of texts, *In*: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, ser. EMSEE '05. Stroudsburg, PA, USA: *Association for Computational Linguistics*, June 30, *p* 13–18.

[2] Rus, V., Lintean, M., Banjade, R., Niraula, N., Stefanescu, D. (2013). Semilar: The semantic similarity toolkit." in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: *Association for Computational Linguistics*, August 4–9, *p* 163–168.

[3] Mihalcea, R., Corley, C., Strapparava, C. (2006). Corpus–based and knowledge–based measures of text semantic similarity, *In*: Proceedings of the 21st *National Conference on Artificial Intelligence* - 1, ser. AAAI '06, vol. 6. Boston, Massachusetts: AAAI Press, July 16–20 2006, *p* 775–780.

[4] Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., Crockett, K. (2008). Sentence similarity based on semantic nets and corpus statistics, *IEEE Transactions on Knowledge and Data Engineering*, 18 (8) 1138–1150, August 2006.

[5] Islam, A., Inkpen, D. (2008). Semantic text similarity using corpusbased word similarity and string similarity," in ACM Transaction on Knowledge Discovery Data, 2 (2). New York, NY, USA: ACM, 2008, *p* 10:1–10:25. [Online]. Available: http://doi.acm.org/10.1145/1376815.1376819

[6] ÇÙ Sari´c, F., GlavaÇÙs, G., Karan, M., ÇÙ, J., Snajder, BaÇÙsi´c, B. D. (2012). Takelab: Systems for measuring semantic text similarity, *In*: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and 2 *In*: Proceedings of the Sixth International Workshop on Semantic Evaluation, ser. SemEval '12. Stroudsburg, PA, USA: *Association for Computational Linguistics*, July 7–8, *p* 441–448.

[7] Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization," Expert Systems with Applications, 36, (4) 7764–7772. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2008.11.022

[8] Steinberger, J., Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation, *In*: Proceedings of 7th International Conference on Information Systems Implementation Modeling, ser. ISIM '04, Ostrava, CZ, April, *p*. 93–100.

[9] Mohler, M., Bunescu, R., Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments, *In*: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-1, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, June 19–24, *p* 752–762. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002568

[10] Rocchio, J. J. (1971). Relevance feedback in information retrieval. Prentice- Hall, Englewood Cliffs NJ.

[11] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *In*: Proceedings of the 5th Annual International Conference on Systems Documentation, ser. SIGDOC '86. New York, NY, USA: ACM, p 24–26. [Online]. Available: http://doi.acm.org/10.1145/318723.318728

[12] Salton, G., Singhal, A., Mitra, M., Buckley, C. (1997). Automatic text structuring and summarization, *Information Processing and Management: an International Journal* - Special issue: methods and tools for the automatic construction of hypertext, 33, (2), p 193–207, March 1997.

[13] Jiang, J. J., Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy, *In*: *Proceedings of International Conference Research on Computational Linguistics, ser. ROCLING* X.

[14] Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets, *IEEE Transactions on Systems, Man and Cybernetics*, 19 (1) 17–30.

[15] Islam, A., Milios, E. E., Keselj, V. (2012). Text similarity using google trigrams, in Advances in Artificial Intelligence: 25th Canadian Conference on Artificial Intelligence, ser. Canadian AI '12. Berlin, Heidelberg: Springer Berlin Heidelberg, May 28–30 2012, p. 312–317.

[16] Kaplan, A. (1995). An experimental study of ambiguity and context, *Mechanical Translation*, 2 (2) 39–46.

[17] Lynum, A., Pakray, P., Gamb ack, B., Jimenez, S. (2014). Ntnu: Measuring semantic similarity with sublexical feature representations and soft cardinality, *In*: Proceedings of the 8th International Workshop on Semantic Evaluation, ser. SemEval '14. Dublin, Ireland: *Association for Computational Linguistics*, August 23–24, p. 448–453.

[18] Malandrakis, N., Iosif, E., Potamianos, A. (2012). Deeppurple: EstiS. mating sentence semantic similarity using n-gram regression models and web snippets, *In*: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, ser. SemEval '12. Montr´eal, Canada: Association for Computational Linguistics, June 7–8 2012, p. 565–570. [Online]. Available: http://dl.acm.org/citation.cfm?id=2387636.2387731

[19] Ramage, D., Rafferty, A. N., Manning, C. D. (2009). Random walks for text semantic similarity, *In*: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ser. TextGraphs–4. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug 07 2009, p. 23–31. [Online]. Available: http://dl.acm.org/citation.cfm?id=1708124.1708131

[20] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy, *In*: Proceedings of the 14th international joint conference on Artificial intelligence - 1, ser. IJCAI '95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., August 20–25, p. 448–453.

[21] Kucera, H., Francis, W. N. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

[22] Resnik, P. (1992). Wordnet and distributional analysis : A class-based approach to lexical discovery, *In*: *A*AAI workshop on statistically-based natural language processing techniques, July, p. 56–64.

[23] Tversky, A. (1977). Features of similarity. psychological review, Tversky, Amos, 84 (4) 327, American Psychological Association.

[24] Sheldon, R. M. (2002). A First Course in Probability, 6th ed. Pearson Education India.

[25] Miller, G. A. (1995). Wordnet: A lexical database for english, *Communications of the ACM*, 38 (11) 39–41.

[26] Finkel, J. R., Grenager, T., Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling, *In*: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, June 25–30, p. 363–370. [Online]. Available: https://doi.org/10.3115/1219840.1219885

[27] Marcus, M. P., Marcinkiewicz, M. A., Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank,*Computational Linguistics - Special issue on using large corpora*: 2 (19) (2), 313–330. [Online]. Available: http://dl.acm.org/citation.cfm?id=972470.972475

[28] Seco, N., Veale, T., Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet, *In*: Proceedings of the 16th European conference on artificial intelligence, ser. ECAI '04. Amsterdam, The Netherlands, The Netherlands: IOS Press, *p*. 1089–1090. [Online]. Available: http://dl.acm.org/citation.cfm?id=3000001.3000272.