Intelligent Information Management, **2013, 5, 171-174**
Published Online November 2013 (http://www.scirp.org/journal/iim)
http://dx.doi.org/10.4236/iim.2013.56018

Semantic Sentence Similarity Using Finite State Machine

Chiranjibi Sitaula[1], Yadav Raj Ojha[2]
[1]Central Department of Computer Science and Information Technology,
Tribhuvan University, Kathmandu, Nepal
[2]Nepal KC Consultancy, Software Company, Kathmandu, Nepal
Email: candsbro@gmail.com

Received September 27, 2013; revised October 25, 2013; accepted November 5, 2013

ABSTRACT

In this paper, a finite state machine approach is followed in order to find the semantic similarity of two sentences. The approach exploits the concept of bi-directional logic along with a semantic ordering approach. The core part of this approach is bi-directional logic of artificial intelligence. The bi-directional logic is implemented using Finite State Machine algorithm with slight modification. For finding the semantic similarity, keyword has played climactic importance. With the help of the keyword approach, it can be found easily at the sentence level according to this algorithm. The algorithm is proposed especially for Nepali texts. With the polarity of the individual keywords, the finite state machine is made and its final state determines its polarity. If two sentences are negatively polarized, they are said to be coherent, otherwise not. Similarly, if two sentences are of a positive nature, they are said to be coherence. For measuring the coherence (similarity), contextual concept is taken into consideration. The semantic approach, in this research, is a totally contextual based method. Two sentences are said to be semantically similar if they bear the same context. The total accuracy obtained in this algorithm is 90.16%.

Keywords: Artificial Intelligence; Natural Language Processing; Text Mining; Semantic Similarity; Finite State Machine

1. Introduction

Semantic similarity is a relationship between common and different features (meaning) of two compared words. It is a fundamental and widely used concept in recent applications of Natural Language Processing. Sentence semantic similarity is main concern to similarity between concepts according to their presumed natural relationships.

In order to compute semantic similarity on the sentence level, we ideally should compare some kind of meaning representation of the sentences. Finding such meaning in Nepali texts is more difficult than English text because the structure matters a lot. The structure used in Nepali sentences is not same as English texts. So, a novel approach is proposed in order to tackle the problem of such texts.

Lots of algorithms and strategies have been developed in natural language processing in order to find the semantic similarity of English texts. But few works have been done in Nepali. And morphologically, processing Nepali text is also of great challenge because it does not match so easily with English sentences. Using the concept of finite state machine, the semantically similar sentences can be retrieved in Nepali texts. The position of subject, verb and object is different from English texts. The concepts of theory of computation and artificial intelligence are major here in this model.

Finite State Automata is an abstract machine which can be in one of the finite number of state. Machine can be in only one state at a time. It can be changed from one state to another by triggering condition for each transition. State machine has been used to describe linguistics —to describe the grammars of natural languages and to learn the corpus patterns.

First Order Logic or Prepositional Logic (also called sentential logic) is a formal language which is used for expressing statements. Prepositional logic is used to determine when a statement is true in a structure. Here, proposition is a statement which is either true or false. Prepositions are connected to give truth or falsity of the compound propositions. Ex: Bi-conditional (or equivalence $\Leftrightarrow$).

Bi-conditional logic is true whenever both of its components have the same truth value, either both true or both false.

Finite State Machine to recognize the bi-conditional logic is as shown in **Figure 1**.

But we can modify it to recognize the single transition state (single word) or double transition states (two words) as shown in the **Figure 2**.

Nepali language has the word order and language writing scripts are different from English language. Nepali language is Subject Object Verb (SOV) language.

2. Literature Review

[1] proposed sentence similarity find method based on Semantic Nets and Corpus Statistics. This method found best for the shortest length sentences. The semantic similarity of two sentences is calculated based on cosine similarity, word order similarity and overall sentence similarity by using information from a structured lexical database and from corpus statistics.

Sentence semantic similarity calculating method based on segmented semantic comparison was proposed in [2]. Best results could be achieved and the calculating process would more fit to the semantic logic in the shortest sentence.

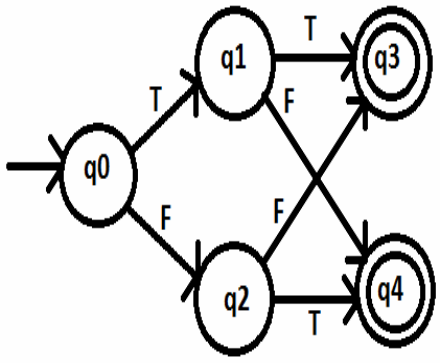

Figure 1. FSA to recognize the bi-conditional logic.

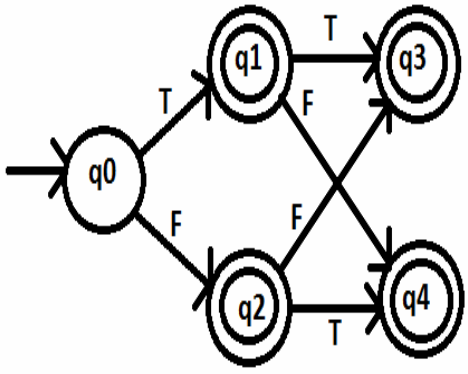

Figure 2. Finite State Machine representation of bi-conditional logic.

[3] described a metric method for computing sentence level semantic textual similarity, which is based on a probabilistic finite state machine model that computes weighted edit distance.

Estimating of the context similarity based on closeness of the semantic load of two comparing sentences is researched by [4].

A FSA is built through a systematic analysis of the patterns of meaning and use for each verb [5]. Also, the algorithm for learning regular grammars from examples to recognize the corpus patterns is given.

Similarly, [6] proposed method based on different aspects like Objects-Specified similarity, Objects-Property similarity, Objects-Behavior similarity and Overall sentence similarity. Sentences would be divided into the trunk and the other segments were set different weights, and the grammatical and semantic structure of the sentences would be analyzed.

The work done by [7] was to produce a measure of semantic similarity, which is a good predictor of "relatedness" between sentences, with the ultimate goal of assessing the coherence of an essay.

Two approaches are proposed by [8], the first approach proposed used lexical similarity & the second used semantic similarity by means of term expansion with synonyms. [9] proposed semantic similarity based on WORDNET dictionary with some sequential process like tokenization, POS tagging, word distance calculation etc.

More importantly, [10] did research on structure of Nepali Grammar. This paper has mainly explained parts of speech of Nepali Language, special characteristics in Nepali Language & overview of the sentential structure of the Nepali Language.

Almost all above methods are based on lexical database WORDNET. Nepali Computational Grammar (NCG) structural detail is done by [11], which essentially involves the development of the intermediate modules like the Parts-of-Speech (POS) Tagger, Chunker and the Parser. [12] built Nepali WORDNET, the rich lexical resource for the Nepali Language for effective machine translation. They did this task inspiration from English WORDNET and Hindi WORDNET. This dictionary can be helpful to get word distance of any two words in the Nepali Language. [13] employed the concept of probabilistic concept for finding the semantic similarity of the sentence.

3. Proposed Model

Almost all existing methods for computing sentence similarity have been adopted from approaches used for long text documents. These methods process sentences in a very high-dimensional space and are inefficient for short text application domain. In order to get accuracy,

computing the similarity between very short texts of sentence length has been proposed. The proposed model considers the first order logic [14] with the finite state machine approach. The research work also is influenced by concept [15].

The semantic concept of contextual [16] is taken into consideration for proposing the new algorithm.

This paper focuses directly on computing the Sentence Semantic Similarity in Nepali Language for short texts. The sample similar sentences are shown in **Tables 1** and **2**.

The similar sentences are those which are contextually similar in this research. For the whole research work, particular context is taken into consideration.

The proposed model is shown in **Figure 3**. It shows the different steps employed for the research activities.

According to **Figure 3**, firstly the document is separated into individual sentence. The individual keyword is taken from each sentence with its semantic orientation. After calculating the semantic orientation of each token, with the help of dictionary, Finite State Machine is made. With the help of finite state machine, the polarity of each sentence is determined. The sentences having same polarity is taken as the same orientation depending on particular context.

4. Evaluation and Output

For the evaluation purpose, around 1200 sentences are taken for the experiment. The datasets used in this experiment are of different type. It was implemented under Visual Studio 2008. Programming language was C#. For measuring the performance, recall is used.

The comparison of the hybrid algorithms with traditional rule based algorithm is made. The output is listed in the below **Table 3**.

The result obtained in **Table 3** showed that the recall of the given algorithm. This algorithm gave different recall values depending of the nature of datasets used. When the datasets of simple nature is used, it gave the recall value of highest *i.e.* 95. Similarly, if the dataset of medium type is used, it gave the recall value of 90.5 and finally with the complex and ambiguous type of sentences, the recall value dropped to 85.

Table 1. First similar sentences.

First Similar Sentences
खिलराज रेग्मि प्रधानमन्त्रि हून्।
उनि प्रधानमन्त्रि होइनन् भन्दिन।

Table 2. Second similar sentences.

Second Similar Sentences
यो घर राम्रो छ।
यो घर नराम्रो होइन।

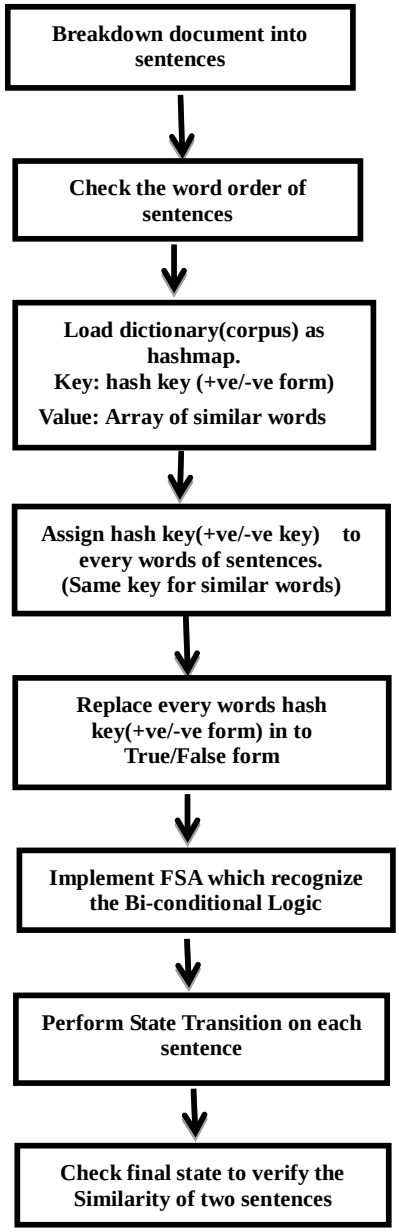

Figure 3. Operational flow of the proposed methodology.

Table 3. Output of proposed algorithm.

Group	# of Sentences	Recall	Average Recall
1	400	90.5	
2	400	85	90.16
3	400	95	

Similarly, the graph is plotted with x-coordinate as the # of sentences with y-coordinate as the recall obtained from the experiment. From the **Figure 4**, it can be seen that the recall value gets increased for simplicity of sentences and has degraded as the complexity of the data set increases. The ambiguity of the sentences degraded the performances.

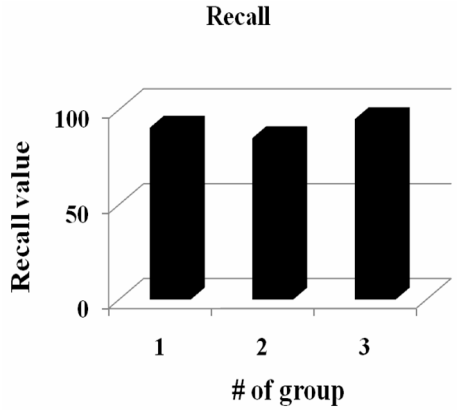

Figure 4. Graph of # of group and recall.

5. Conclusion and Limitation

The implemented algorithm gave recall of 90.15%. The algorithm is the exploitation of the finite state machine with bi-directional logic. The bi-directional logic is modified here in order to perform the semantic similarity of the sentences. The semantic ordering works have to be performed carefully otherwise the performance may be degraded. The sentences used in this experiment are context based. It can be compared with uni-directional logic of artificial intelligence for further works. Although the algorithm is designed for Nepali texts, it can also be used in other languages.

6. Acknowledgements

I would like to thank Prof. Dr. Sashidhar Ram Joshi from IOE, Tribhuvan University, Nepal for providing me implausible support.

REFERENCES

[1] Y. H. Li, *et al.*, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 8, 2006, pp. 1138-1150. http://dx.doi.org/10.1109/TKDE.2006.130

[2] Y. T. Liu and Y. J. Liang, "A Sentence Semantic Similarity Calculating Method Based on Segmented Semantic-Comparision," *Journal of Theoretical and Applied Information Technology*, Vol. 48, No. 1, 2013, pp. 231-235.

[3] M. Q. Wang and D. Cer, "Stanford: Probabilistic Edit Distance Metrics for STS," Unpublished.

[4] M. Mehdi and S. M. Fakhrahmad, "Effective Estimation of Context Similarity: A Proposed Matching Model Based on Weighted Semantic Load," *International Journal of Artificial Intelligence & Applications*, Vol. 3, No. 3, 2012, pp 1-10.

[5] O. Popescu, "Learning Corpus Patterns Using Finite State Automata," FBK-irst, Trento, 2013.

[6] L. Li, *et al.*, "Measuring Sentence Similarity from Different Aspects," *Proceeding of the* 8*th International Conference on Machine Learning and Cybernetics*, Baoding, 12-15 July 2009, pp. 2244-2248.

[7] D. Higgins and J. Burstein, "Sentence Similarity Measures for Essay Coherence," *Proceedings of the* 7*th International Workshop on Computational Semantics* (*IWCS*), Tilburg, 2007, pp. 1-12.

[8] D. Vilarino, *et al.*, "BUAP: Lexical and Semantic Similarity for Cross-Lingual Textual Entailment," *Proceedings of the* 1*st Joint Conference on Lexical and Computational Semantics*, Montreal, 7-8 June 2012, pp. 706-709.

[9] T. N. Dao and T. Simpson, "Measuring Similarity between Sentences," Unpublished.

[10] B. K. Bal, "Structure of Nepali Grammar," PAN Localization, Madan Puraskar Pustakalaya, Kathmandu, Nepal, 2004, pp. 332-396.

[11] P. Rupakheti, L. P. Khatiwada and B. K. Bal, "Report on Nepali Computational Grammar," Unpublished, pp. 1-25.

[12] A. Chakrabarty, B. Purkayastha and A. Roy, "Experiences in Building the Nepali Wordnet-Insights and Challenges," *The* 5*th Global Wordnet Conference at CFILT, IIT Bombay*, Mumbai.

[13] I. Beltagy, *et al.*, "Montague Meets Markov: Deep Semantics with Probabilistic Logical Form," 2*nd Joint Conference on Lexical and Computational Semantics*: *Proceeding of the Main Conference and the Shared Task*, Atlanta, 13-14 June 2013, pp. 11-21.

[14] S. Ferilli, *et al.*, "Plugging Taxonomic Similarity in First-Order Logic Horn Clauses Comparison," *AI*IA* 2009: *Emergent Perspectives in Artificial Intelligence*, *Lecture Notes in Computer Science*, Vol. 5883, 2009, pp. 131-140. http://dx.doi.org/10.1007/978-3-642-10291-2_14

[15] D. K. Lin, "An Information-Theoretic Definition of Similarity," *ICML*, Vol. 98, 1998, pp. 296-304.

[16] C. Sitaula, "Semantic Text Clustering Using Enhanced Vector Space Model using Nepali Language," *GESJ*, Vol. 36, No. 4, 2012, pp. 41-46.