Paraphrase Identification Using Textual Entailment Recognition

Mrs. Seethamol S
Department of Computer Science and Engineering
College of Engineering Cherthala
Kerala, India-688541
Email: seetha.santhan486@gmail.com

Mrs. Manju K
Department of Computer Science and Engineering
College of Engineering Cherthala
Kerala, India-688541
Email: manju@cectl.ac.in

***Abstract*—Paraphrase Identification is a challenging and popular research area in Natural Language Processing Applications, including document summarization, information retrieval, question answering etc. In this paper, we focus on paraphrase identification using textual entailment recognition. In the proposed method Paraphrasing is seen as a bidirectional textual entailment. The predicate argument structures are used to represent the text and distant supervision technique for mining the indirect facts from these texts. The proposed method uses a probabilistic network for judging the confidence of each fact and a word sense disambiguation module is used to check the similarity between two sentences.**

***Keywords*—Paraphrase Identification ; Textual Entailment ; Word Sense Disambiguation ; Distant Supervision.**

I. Introduction

Paraphrase Identification and Textual entailment are two different yet well related tasks in Natural Language Processing. Paraphrase Identification is a classification task. Given a pair of sentences, the system should examine the two sentences and determine whether they have the same meaning or not. On the other hand, textual entailment in natural language processing is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text.

For example,

T: The disaster happened near the Greek bank.
H: The explosion happened near the building.

Here Text T textually entails Hypothesis H when the bank in T is referred as a financial institution, not in the sense of a river. ie, for paraphrase identification as bidirectional entailment, the context of words in the text hypothesis pair should be considered. To solve this problem, similarity between the text hypothesis pairs are calculated by performing Word Sense Disambiguation and textual entailment recognition is also performed, so that a human who reads T would infer that H is most likely true.

This paper is organized as follows: section II briefly discusses some of the methods that have been used for Textual Entailment Recognition. The proposed system for textual entailment recognition is explained in the section III, followed by the result analysis and discussion in section IV. Finally section V concludes the paper and discusses the future works.

II. Related Work

PASCAL Recognizing Textual Entailment (RTE) Challenges [2] have introduced a collection of systems which can be used to recognize the textual entailment pairs. Approaches like logical, probabilistic, graph matching and machine learning are mainly used in these systems.

In 2014 Yann Huei Lee et al.[11] introduced an approach which work directly on the input surface strings. Here the preprocessing techniques like part of speech tagging or named entity recognition are used. The limitation of these approach is, the entailment decision is calculated by only considering the lexical evidences and with out retrieving the syntactic or semantic details from the text.

Rohini Basker et al.[6] takes textual entailment as a graph matching problem. But the texts and hypothesis are converted into trees or graphs. Once these graphs or trees are obtained, the matching score between these graphs are determined and if the score achieves the threshold value then the entailment is considered as true. Nidhi Sharma et al.[12] build a machine learning system which uses the similarity features and consider the textual entailment as a classification task.

Recently, Lei Sha et al.[13] proposed the concept of predicate-argument structure to represent text and hypothesis. Using YAGO database they performed the distant supervised learning technique to mine indirect facts for the selected predicates. Then using these facts construct the probabilistic network and calculated the probability of each facts with the help of Markov Logic Network (MLN)[4]. This probability value is further used for textual entailment recognition.

RTE is the task of recognizing whether the information expressed in a Hypothesis can be inferred from the Text. Considering the above, predicate argument structure method can yield better bidirectional textual entailment results. However, the limitation of Lei Sha et al.[13]'s work is, we cannot tell if two given sentences are paraphrases or not without considering the context in which the words are appeared.

III. System Overview

This paper discusses an RTE framework for paraphrase identification. An overview of the system is shown in figure

978-1-5090-6106-8/17/$31.00 ©2017 IEEE

1. The role of each module in figure 1 is further explained in detail. An algorithm is also designed to illustrates how the RTE system works and the overall system can be explained using algorithm 1.

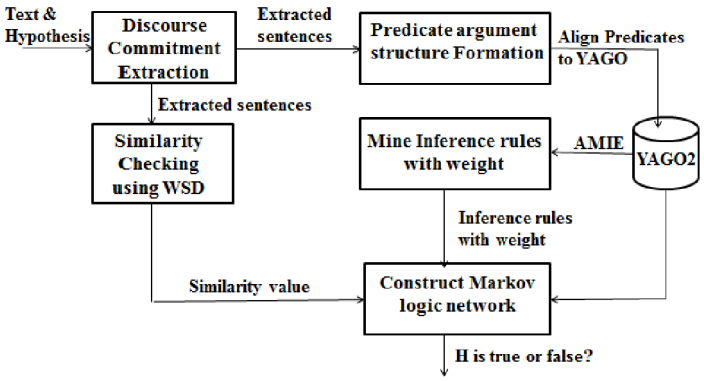

Fig. 1. Textual Entailment Recognition System

Algorithm 1 Textual Entailment Recognition:

Require: Text Hypothesis sentence pair.
Ensure: Text entailment decision.

1: Convert the Text and Hypothesis into discourse commitments.
2: Convert the discourse commitments into Predicate argument structure.
3: Align the predicates into the YAGO database.
4: AMIE system is used to mine the weighted inference rules from YAGO facts.
5: Calculate the similarity between Text Hypothesis pairs and obtain the similarity value.
6: Construct Markov logic network using the similarity value and inference rules for the predicate in Text and find the probability of each facts.
7: **If**the predicate in text with high probability is same as the predicate in hypothesis, then return the entailment decision as true.
8: **else** return the entailment decision as false.

A. Discourse Commitment Extraction

Discourse commitment system[7] decomposes the text sentence to a sequence of precise and understanding sentences which convey the exact meaning of original sentence. The main advantage of using discourse commitment is that it uses both syntactic and semantic rules to deduce the unexpressed facts of the text sentences. Discourse commitment extraction is performed using the algorithm generated by Hickl.

B. Predicate Argument Structure Formation

Once the discourse commitments are extracted, transfer all these extracted sentences into predicates. REVERB[9] is used to extract the discourse commitments in the form of (predicate + 2 Arguments) triples. REVERB is an open extractor which automatically identifies and extract the binary relations from a sentence. A confidence value is assigned to the extracted relations using logistic regression classifier. In YAGO database facts are represented in predicate argument format. Therefore, in predicate extraction only the noun phases are considered as arguments.

C. Distant supervision with YAGO (Predicates Alignment)

YAGO[14] is a very large database which represent the facts in predicate argument format. In this database every facts are connected with each other. Therefore, the predicates in the T-H pair are aligned to the YAGO facts for using these connections[8]. Since YAGO is very large, the common predicates can easily mapped to YAGO facts. If YAGO did not match the predicate in T then check the similar predicates as its synonymous predicates. If the original predicate are not found in YAGO, then consider the 10 almost identical predicates. Still no match is found then it is concluded that the given predicate has very less connection with the other predicates and cannot be supervised by YAGO.

D. Similarity Checking using Word Sense Disambiguation

The goal of Similarity checking using Word Sense Disambiguation is used for the purpose of implementing the paraphrase identification using textual entailment recognition. Here check the similarity between text and hypothesis sentences and explicit the correct meaning of the expressions due to word senses and syntactic ambiguities. Here word sense disambiguation is performed with the help of natural language tool kit(nltk).

E. Probabilistic Inference

Markov Logic Network [4] is used to implement the probabilistic inference in the RTE system. MLN is constructed using similarity value and inference rules where every rule has a weight assigned according to the real word facts. Here AMIE [10] system is used to extract inference rules from YAGO. AMIE is a YAGO based inference rule mining system where each rule is associated with a confidence value.

In probabilistic inference [13], relational information in YAGO are given to the Markov logic network and inference rules with weights are used for training purpose. Given the predicates in T, first construct a Markov logic network by selecting the related rules. MLN then compute the unknown fact probabilities which used for entailment decision. For example, given the facts as motherOf (Hillary, Chelsea) and marriedTo(Hillary, Clinton), Markov logic network will calculate all the possible predicates like motherOf(Hillary, Clinton), fatherOf(Clinton, Chelsea), marriedTo(Clinton, Chelsea) etc. The fact which is most likely to be true will have highest probability. ie, fatherOf(Clinton, Chelsea) in the above example.

IV. Experiments and Results

For conducting the experiments, first evaluate the performance of proposed RTE framework on the PASCAL RTE-2[3] and RTE-3[5] dataset, which contain 1600 examples. REVERB[9] is used to generate the predicate argument structures in RTE system. Then use the YAGO2 for aligning predicates and mining inference rules. YAGO2 is a large

database with more than 900K facts and 400K entities. Then run the AMIE system[10] on YAGO2 for only one time to get all inference rules. For each T-H pair, here only select a small portion of related inference rules to construct Markov logic network. The chosen rules should contain not less than one predicate in the predicates obtained from T-H pair. In addition here use the natural language tool kit for checking the similarity between text and hypothesis pairs. This will consider the exact meaning behind each text and increases the accuracy of entailment recognition.

TABLE I. PERFORMANCE ANALYSIS OF PROPOSED METHOD WITH CURRENT RTE METHODS

Approach	Accuracy
Lexical Approaches	67.50%
Graph Based Approaches	65.33%
Machine Learning Approaches	77.00%
Discourse Commitment Approaches	84.93%
Predicate logic + similarity Approach	85.93%

Here compared the results with 4 standard systems: (1) Yann-Huei Lee et al. (2014)s Lexical approach, (2) Rohini Basker et al. (2007)s graph-matching approach, (3) Nidhi Sharma et al. (2015)'s Machine Learning approach. (4) Hickl (2008)s discourse commitment based method. The comparison of these 4 systems and proposed framework is shown in Table I. Since in textual entailment we only needed to determine whether the entailment is positive or not for the text hypothesis pair. ie, the precision is equal to the recall, so here represent the precision only. According to the Table I, the performance of proposed framework is higher than others.

Definition of a paraphrase would insist on the two sentences which have strictly identical meanings. ie, every details of syntactic and lexical changes are considered. But to captivate more interesting results authors of Microsoft Research Paraphrase Corpus(MSRPC)[1] relaxes the definition of paraphrase to almost identical meaning. In MSRPC sentences are considered paraphrases if it is mostly bidirectional instead of fully bidirectional[1]. The corpus consist of a large set of sentence pairs S1; S2, together with labels indicating whether the sentences are in a paraphrase relationship or not.

The system explained in section III is tested against the Microsoft Research Paraphrase Corpus. All WordNet word senses were examined when calculating the similarity between words in the sentences. For each similarity metric MSRPC is used for the training purpose which will classify the threshold for the similarity value and increase the accuracy. Here used 870 training and 720 test examples for this work.

The performance analysis and comparison of different systems are done using Precision and Recall. For analysis, here generated a training and test set using from MSRPC. The training and test data consists of 1600 examples. For the first experimentation, here used predicate logic for bidirectional textual entailment. In this the harmonic mean of precision and recall is Fmeasure = 0.851 and in second experiment the String Similarity approach is used. In this the Fmeasure = 0.716. Finally, the third experiment used proposed method which combines both predicate logic and String Similarity approach. This system achieves Fmeasure = 0.863. The results are shown in Table II. Here the proposed method have greater accuracy and efficiency than other 2 methods.

TABLE II. PERFORMANCE ANALYSIS OF PARAPHRASE IDENTIFICATION

Approach	Accuracy	F-measure
Predicate logic Model	78.1%	85.1%
String similarity model	66.10%	71.6%
Predicate logic + similarity Approach	78.6%	86.3%

V. CONCLUSION

In this paper discussed an approach for the task of paraphrase identification using textual entailment recognition. Here, Paraphrasing is seen as a bidirectional textual entailment. In the proposed RTE system the predicate-argument structures is used to represent the text hypothesis pair and used the distant supervised learning technique to mine indirect facts for the selected predicates. The system also used a probabilistic network for judging the confidence of each facts in RTE and a word sense similarity checking to explicit the more accurate results for paraphrase identification. Comparing with the existing systems this framework may accurately recognize the entailment pairs and Identify the paraphrase sentences. In future, we can extend this framework into document summarization systems.

ACKNOWLEDGMENT

We would like to thank those who supported for the completion of our work.

REFERENCES

[1] William B. Dolan and Chris Brockett. 2005. "Automatically constructing a corpus of sentential paraphrases", *In The 3rd International Workshop on Paraphrasing* (IWP2005).

[2] Ido Dagan et al. 2006. "The PASCAL Recognizing Textual Entailment Challenge", *MLCW* , LNAI Volume 3944, pages 177-190. Springer-Verlag.

[3] Roy Bar-Haim et al. 2006. "The Second PASCAL Recognizing Textual Entailment Challenge", *In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, Venice, Italy.

[4] Matthew Richardson and Pedro Domingos. 2006. "Markov logic networks", *Machine learning*, 62(1- 2):107-136.

[5] Danilo Giampiccolo et al. 2007. "The Third PASCAL Recogniz-ing Textual Entailment Challenge", *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

[6] Rohini Basker et al. 2007. "Recognizing Textual Entailment by Soft Dependency Tree Matching" in , *ISSN 2007-9737*

[7] Andrew Hickl. 2008. "Using discourse commitments to recognize textual entailment", *In Proceedings of the 22nd International Conference on Computational Linguistics* Volume 1, pages 337344. Association for Computational Linguistics.

[8] Mike Mintz et al. 2009. "Distant supervision for relation extraction without labeled data", *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Volume 2-Volume 2, pages 10031011.

[9] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. "Identifying relations for open information extraction", *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 15351545. Association for Computational Linguistics.

[10] Luis Antonio Galarraga et al. 2013. "Amie: association rule mining under incomplete evidence in ontological knowledge bases", *In Proceedings of the 22nd international conference on World Wide Web*, pages 413422.

[11] Yann-Huei Lee et al. 2014. "Recognizing Textual Entailment using Lexical, Syntactical, and Semantic Information", *In Proceedings of the 11th NTCIR Conference*, December 9-12, 2014, Tokyo, Japan.

[12] Nidhi Sharma et al. 2015. "Recognizing Textual Entailment using Dependency Analysis and Machine Learning", *In Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW)*, pages 147153, Denver, Colorado, June 1, 2015. c 2015 Association for Computational Linguistics.

[13] Lei Sha, Sujian Li et al. 2015. "Recognizing Textual Entailment Using Probabilistic Inference", *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 16201625, Lisbon, Portugal, 17-21 September 2015.

[14] Farzaneh Mahdisoltani, Joanna Biega et al. "Yago3: A knowledge base from multilingual wikipedias", *In 7th Biennial Conference on Innovative Data Systems Research*. CIDR 2015.