DOI: 10.7763/IPEDR. 2013. V62. 15

A Novel Graph-Based Recognizing Textual Entailment System

Elaheh Hosseini [1+], Hossein Sameti [1]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Abstract. A graph-based Recognizing Textual Entailment (RTE) system is proposed to first, study the effect of statistical features on RTE system, second, cut the cost of graph alignment by introducing a novel graph-based RTE system that uses just one simple graph instead of two. A simple graph is applied to text representation and a Conditional Random Fields (CRF) is used to measure similarity and decision making. The firs is construct by a function of PMI while the latter is learned based on statistical features extracted from simple graph to recognize the existence of entailment. An edit distance algorithm is applied as a decision maker in the second level of the system on low confidence samples to increase the confidence of system performance. The system is trained on the RTE competitions development sets RTE1, RTE2, and RTE3 and tested on the respective RTE test sets. Accuracy is used as evaluation measure to compare the results with some other RTE complicated lexical-based systems.

Keywords: Recognizing Textual Entailment; Graph Representation; Linear Chain CRF; Statistical Dependency.

1. Introduction

One of the recently defined challenges in the field of natural language understanding is Recognizing Textual Entailment (RTE). It has been studied in different applications and from different aspects of language processing. For the first time Glickman and Dagan introduced the RTE. According to their definition text T entails hypothesis H; if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T [1]. After their work, many different definitions of RTE have been introduced based on application, processing level, and types of data. Some of these innovative algorithms are as follows: logical entailment [2], machine translation evaluation methods [3], RTE based on web search and entailment rules [4], graph based entailment recognizing systems [5].

Processes perform at different levels, could be studied in different aspects. Usually lexical level inferences are simple [6][7]. In addition to these systems, some of the upper level entailment systems estimate the similarity of Text and Hypothesis by making decision on the words of document [8]. The methods introduced at sentence processing level rely on deep analysis of World Knowledge or perform based on shallow analysis like simple graph matching algorithms [9]. Graph based RTE algorithms, have been studied for a long time and include many different methods such as entailment graphs, syntactic graphs, and semantic graphs [9][10][11]. We recommend a combined entailment method based on a graphical model. At first a word entailment engine based on statistical features is produced by applying a simple graph and conditional random fields. A word similarity recognizing entailment based on edit distance algorithm is used to complete the system in low confidence condition.

The rest of the paper is organized as follows: Section 2 describes the system structure and details of processing level and applied features. The simulation and results are reported in section 3. Section Finally, we draw the conclusions in Section 4.

2. System Structure

This part of the model attempts to distinguish relations between Hypothesis terms and the Text Document terms based on graphical modelling. Figure 1 shows the components of the proposed system. First of all, the problematic symbols in parsing are chosen manually and listed as noisy symbols, and then all of them are replaced with blank. After noise reduction, we tokenized and lemmatized the corpus by using

[+] Elaheh Hosseini. Tel:+989385247301; +98 (21) 66019246
elaheh_hosseini@ce.sharif.edu, hoseini_1377@yahoo.com

Stanford open source tools. We use undirected graphs as a classifier. Textual entailment is a directional process in which Hypothesis must be entailed by the Text Document. We suppose several sub-assumptions in extracting graph-based features. First, as the Hypothesis terms must be entailed so some of the Text terms may not be applicable in this process. Second, Synonyms, usually can take the same syntactic roles in the sentence so they are expected to show the same semantic relation with their neighbour words in the sentence. Third, usually the meaningful probabilistic relations decrease as the distances of words in the context increase, so for simplification and efficient calculation we use a window with a proper size to consider just the most efficient relations. At low confidence situation, we construct a token edit distance without considering any syntactic or semantic feature such as POS or synonyms. As the Text document must entail the Hypothesis and not vice versa, so the words of the Hypothesis are taken into consideration rather than the Text document. This could be modelled by zero cost for deletion operation. Conditional random field (CRF) is applied as a classifier in order to make the first decision on entailment system. Confidence measure of this level is calculated by the CRF output probabilistic. Then a token edit distance is used to address words relation and sentence structure. For confidence measure we test several statistical and non statistical measures such as Information, PMI, cosine similarity or Jaccard coefficient. The voting module works based on confidence measure of both results.

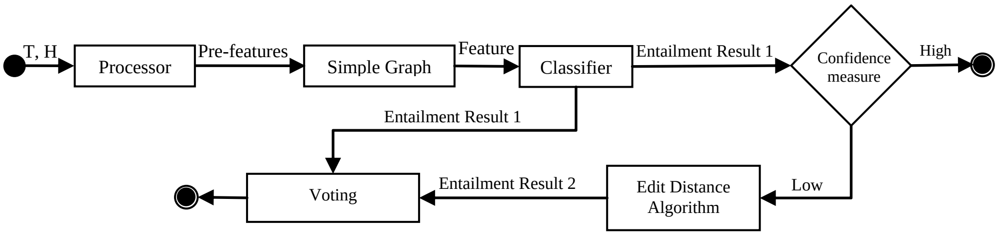

Fig. 1: Structure of the proposed textual entailment system

2.1. Linear Chain Conditional Random Field Classifier

Conditional random fields have been used in textual entailment task as a discriminative learning classifier to implement tree edit distance algorithm by Wang and Manning on sentence processing level [12]. In this paper we used CRF as a classifier to implement our RTE engine. Conditional random fields are successfully applied in several different tasks in natural language processing such as sequence labelling, segmentation and classification. Given an input data $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ and an output labels $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$, the linear chain conditional random field model is as follows;

$$P_\lambda(y\,|\,x) = \frac{1}{Z(x)} exp(\sum_{k=1}^{K} \lambda_k f_k(y,x)).$$

In this equation $\boldsymbol{Z}(\boldsymbol{x})$ roles as normalization factor which is a function of $\boldsymbol{x}$. The feature functions $\boldsymbol{f_k}(\boldsymbol{y}, \boldsymbol{x})$ are defined as $\boldsymbol{f}_{\boldsymbol{y}',\boldsymbol{j}}(\boldsymbol{y}, \boldsymbol{x}) = 1_{\{\boldsymbol{y}'=\boldsymbol{y}\}} \boldsymbol{x_j}$ for all feature weight and for bias weight, the feature function is $f_{y'}(y,x) = 1_{\{y'=y\}}$ Weights of feature functions $\boldsymbol{\lambda_k}$ are considered as model parameters. In classification tasks the new data label will be defined through dynamic programming methods such as iterative forward-backward estimation, to perform suitable parameter estimation. The hypothesis terms play the role of input data, "entails", and "contradicts" are the labels used in the output of this part of the model. Feature functions are defined according to Hypothesis and Text document. The entailment of Hypothesis will be decided based on amount of occurred changes in the Text Document graph by adding Hypothesis information to it.

2.2. Statistical Graphical Modelling

In natural language processing a vast range of features at different language processing levels including the simplest features like IDF and some of the most complicated semantic features are applied. In this paper, we try to use the most possible simple features such as frequency-based features to study the effects of statistical relation of document components on understanding the meaning and RTE. The simple graph of Hypothesis $G_H(V,E)$ is defined as follows: vertexes $v \in V$ are the document words, edges $e \in E$ show the relations of words. The weight of each link is defined as the Pointwise Mutual Information (PMI) between words so the links of the graph are directed connection between nodes. The feature which measures the degree of dependency between two words is defined by Church and Hanks in 1990 for the first time [13] as $PMI_{(W_H,W_T)} = log_2 \frac{P(W_H,W_T)}{P(W_H)P(W_T)}$ It is calculated at sentence level and all probabilities are counted by the number of sentences that contain the mentioned words. The formula is modified to calculate it at the word level. In other words, we count the bi-gram, so we won't count the sentences which contain both W_H and W_T words. Instead if the W_H and W_T occur more than once in a sentence it will be counted correspondingly. So in the formula $P(W_H,W_T)$ stands for the probability of W_H and W_T occurrence in the document. $P(W_H)$ and $P(W_T)$ are the probabilities of W_H and W_T occurrence in these documents respectively. We used PMI as a statistical feature defined by the primary relation between context words. This feature defines statistical relations of both states and state-observations in our undirected graphical model. Weight of simple graph is defined as follows: $w_{i,j} = \frac{c + r(PMI(i,j))}{d_{i,j}^2}$ In this equation r stands for the ramp function, C is constant, $\boldsymbol{d_{i,j}^2}$ shows the distance between two words in the document. For T-H links we define the distance by subtracting the words' positions in Text document and Hypothesis. Since it is assumed, increasing the words' distance decreases the statistical relations between them, so an about ten-word distance window is used to prevent meaningless calculation. Text document graph; $G_T(V,E)$ is constructed in the same way as $G_H(V,E)$. To extracting dependency relation features between Hypothesis and Text document, another simple graph $G_{HT}(V,E)$ is defined. The simple graph's features used in this paper are:

- Number of common neighbors for each uncommon vertex in Hypothesis graph. Common vertexes are those exist in both Hypothesis graph and Text document graph.
- Number of all $\boldsymbol{G_H}$ nodes' neighbors among uncommon T nodes.
- Normalized sum of weights for each nodes in $\boldsymbol{G_H}$ graph
- Normalized sum of weights for each nodes in $\boldsymbol{G_{HT}}$ graph
- Ratio of normalized neighbor numbers of $\boldsymbol{G_H}$ graph to that of T graph. Normalization is based on documents' length.
- Ratio of connectivity of $\boldsymbol{G_H}$ graph to that of $\boldsymbol{G_{HT}}$ graph
- Ratio of total weights of each node in $\boldsymbol{G_H}$ graph to that of $\boldsymbol{G_T}$ graph
- Normalized number of all $\boldsymbol{G_T}$ graph nodes to the connectivity of $\boldsymbol{G_H}$ graph
- Normalized ratio of connectivity of $\boldsymbol{G_H}$ graph to the connectivity of $\boldsymbol{G_T}$ graph
- Ratio between Number of adding links to $\boldsymbol{G_T}$ graph to the sum of weights of new links to $\boldsymbol{G_T}$ graph with new nodes of $\boldsymbol{G_H}$ graph
- • Normalized number of neighbors of each vertex in $\boldsymbol{G_H}$ graph
- • Ratio of normalized number of connecting edges in $\boldsymbol{G_H}$ graph to that of $\boldsymbol{G_T}$ graph

3. Simulations and Results

We used three standard RTE data sets RTE1 and RTE3. In this section the experiments and their results are explained in detail. We compared our system with some lexicon-based algorithms [14], which use WordNet similarity and dependency based similarity measures [15]. We have represented our document according to its statistical relations by conditional random field. Statistical features are extracted with the same strategy as mentioned in 2.2 introduced based on [16] for simple graph text representation method. The achieved results of first level will be evaluated according to confidence measure. In this position, string edit distance algorithm is applied to low confidence pairs. To show the effect of statistical features, syntactic or semantic features are not considered in this level. Comparing the proposed system with other ones shows its acceptable performance that stems from its hierarchical structure. Tables 1 and 2, show the results of our system on two RTE datasets. We chose accuracy to report our achieved results. For RTE1, We used the first development set in training. We compared our system with other similar systems based on lexicon entailment.

Table 1: Achieved accuracy on RTE1 test set by different systems

Lexical Entailment System Description	Accuracy %
Proposed system	49.88
Proposed system with Syntactic features	50.76
Wu (HKUST): Statistical lexical relations and Syntactic matching	51.20
Andreevskaia and et al (Concordia): WordNet and Syntactic matching	51.90
Zanzotto (Rome-Milan): WordNet and Syntactic matching	52.40
Jadavpur System: Lexical Relation, WordNet and Syntactic matching	53.70

Table 1 shows first RTE challenge participated systems' results using different word relations for recognizing entailment in text. Wu used statistical and lexical relations and considered two possible, Left-to-Right and Right-to-Left, orders for the words appeared in the text [17]. Concordia and Rome universities used syntactic and semantic relations. Jadavpur System used different lexicon, syntactic and semantic relations. They use unigram, bi-gram, longest common sub-sequence, syntactic relation such as Subject-Verb or Object-Verb relation WordNet based semantic relation. We referred to the results of other similar systems mentioned in [8]. In this system the suitable feature combination and low-level result usage enhance the performance of system. Reported accuracy in all cases of our systems relate to the best condition of training data and the worst condition of test data. Experiments show that accuracy of our system is highly related to the task. The accuracy decreased in case of the tasks highly changes based on the structure of the sentence.

Table 2: Achieved accuracy on RTE3 test set by different systems

Lexical Entailment System Description	Accuracy %
Proposed system	57.50
Proposed system with Syntactic features	59.29
Ha Harmling: Lexical Relation, WordNet, Syntactic, Matching /Aligning, ML Classification	57.75
Blake: Lexical Relation, WordNet, Syntactic Matching/Aligning, ML Classification	60.50
Ferrés: Lexical Relation, WordNet, Syntactic Matching/Aligning, ML Classification	60.62

3.1. System Results in Different Applications

In this experiment, the data samples are separated based on tasks. The number of samples is not the same in different tasks. Also the number of similar words in each sample pairs of document, sentence structure similarity between Text and Hypothesis and length of documents in each sample pairs are not the same. Achieved results show not only performance of the proposed system is depend on statistical features but also it depends on task, number of training sample pairs, structure of sentence, and word similarity.

Table 3: Percent of accuracy achieved by proposed system in training

RTE1_DEV1	79.79
RTE1_DEV2	76.07
RTE2	82.13
RTE3	83.50

Table 4: Percent of accuracy achieved by proposed system in Test for RTE Datasets

Application	% Accuracy			
	RTE1_DEV1	RTE1_DEV2	RTE2	RTE3
IE	55.00	50.83	43.0	50.5
IR	61.11	64.44	49.0	57.99
RC	45.71	45.0	*	*
CD	58.67	66.67	*	*
PP	46.0	56.00	*	*
MT	45.83	45.0	*	*
QA	47.69	40.77	61.5	58.5

Table 4 shows the achieved results of the proposed system for RTE1, RTE2, and RTE3 datasets. The difference in the results of the same tasks for RTE1 is caused by the number of sample pairs in training and difference in these samples. The results of the system for RTE2 and RTE3 datasets show that, although the number of sample pairs for all tasks are equal, the effect of features and document structure in different tasks cause noticeable differences on the achieved results. The difference in the same tasks in RTE2 and RTE3 are high but the results in training sets in table 3 do not change a lot for these datasets. It could be concluded that the RTE3 test set is more similar to RTE3 train set in the data format than RTE2 test and train sets.

4. Conclusion and Future Works

Usually, RTE systems apply sophisticated semantic or syntactic features or utilize complex methods like logical rule, knowledge based engines or web mining engines. We constructed our system based on the simple statistical features and attempted to enhance its performance by a hierarchical strategy and bi-step process. By using the probability of entailment occurrence as confidence measure and processing the low confidence pair, the accuracy of our system is increased. As we constructed just one graph in our system, so graph alignment and its problems will be avoided.

The results of proposed method on three datasets of the RTE tracks is reported and proved acceptable performances of our system. The results show that the main challenges of our system are varieties in syntactic structures of sentence and using lots of synonyms. To overcome these weak points we plan to add syntactic and semantic features to our system. One drawback of our system is determining the accurate distance between Hypothesis terms and Text terms. As a future work we are going to implement locally adaptable edit operations to provide more accurate decision criteria.

5. References

[1] Dagan, I. & Glickman, O., Probabilistic textual entailment: Generic applied modeling of language variability. *PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble*, 2004.

[2] Roth, D. & Sammons, M., Semantic and logical inference model for textual entailment, *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007, pp. 107-112.

[3] Neogi S, Pakray P, Bandyopadhyay S, and Gelbukh A., JU-CSE-NLP: Language Independent Cross-lingual Textual Entailment System. *Proc. of the 6th International Workshop on Semantic Evaluation (SemEval 2012),* 2012, pp. 689-695.

[4] Szpektor, I. & Dagan, I., Learning Entailment Rules for Unary Templates, *Proc. of the 22nd International Conference on Computational Linguistics (COLING),* 2008, pp. 849-856.

[5] Berant, J., Dagan, I. & Goldberger, J., Global learning of focused entailment graphs, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1220-1229.

[6] Shnarch E., Goldberger J., Dagan I., Towards a probabilistic model for lexical entailment, *Proc. of the TextInfer 2011 Workshop on Textual Entailment (TIWTE)*, 2011, pp. 10-19.

[7] Adams, R., Nicolae, G., Nicolae, C. & Harabagiu, S., Textual entailment through extended lexical overlap and lexico-semantic matching, *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007, pp. 119-124.

[8] Pakray, P., Gelbukh, A., & Bandyopadhyay, S., A Syntactic Textual Entailment System Using Dependency Parser, Springer, Vol. 6008, *Book Computational Linguistics and Intelligent Text Processing*, 2010, pp. 269-278.

[9] Heilman M & Smith N., Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. *Proc. of Human Language Technologies (HLT): The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL),* 2010, pp. 1011-1019.

[10] Berant J, Dagan I, and Goldberger J., Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 2012, 38(1): 73–111.

[11] Androutsopoulos, I. & Malakasiotis, P., A survey of paraphrasing and textual entailment methods, *Artificial Intelligence Research*, 2010, 38(1): 135-187.

[12] Wang, M., C., D., Manning, Probabilistic Tree-Edit Models with Structured Latent Variables for Textual Entailment and Question Answering, *COLING '10, Proc. of the 23rd International Conference on Computational Linguistics,* 2010, pp. 1164-1172.

[13] Church, K. W. & Hanks, P. Word association norms, mutual information, and lexicography, *Computational linguistics*, 1990, 16(1): 22-29.

[14] Hirst, G. & St-Onge, D., Lexical chains as representations of context for the detection and correction of malapropisms, *In WordNet: An Electronic Lexical Database, ed., Christiane Fellbaum, c*hapter 13, 1998, pp. 305–332.

[15] Lin, D., An information-theoretic definition of similarity, *Proc. of the Fifteenth International Conference on Machine Learning (ICML)*, 1998, pp. 296-304.

[16] Gamon, M., Graph-based text representation for novelty detection, *Association for Computational Linguistics, TextGraphs-1, Proc. of the First Workshop on Graph Based Methods for Natural Language Processing*, 2006, pp. 17-24.

[17] Wu, D., Textual Entailment Recognition Based on Inversion Transduction Grammars, *The PASCAL Recognising Textual Entailment Challenge, Proc. of the First PASCAL Recognizing Textual Entailment (RTE) Workshop*, 2005.