Recognizing Textual Entailment with Statistical Methods

Miguel Angel Ríos Gaona[1], Alexander Gelbukh[1], and Sivaji Bandyopadhyay[2]

[1] Center for Computing Research, National Polytechnic Institute, Mexico
mriosb08@sagitario.cic.ipn.mx, gelbukh@gelbukh.com
[2] Computer Science & Engineering Department, Jadavpur University, Kolkata 700 032 India
sivaji_cse_ju@yahoo.com

Abstract. In this paper we propose a new cause-effect non-symmetric measure applied to the task of Recognizing Textual Entailment .First we searched over a big corpus for sentences which contains the discourse marker "because" and collected cause-effect pairs. The entailment recognition is based on measure the cause-effect relation between the text and the hypothesis using the relative frequencies of words from the cause-effect pairs. Our measure outperformed the baseline method, over the three test sets of the PASCAL Recognizing Textual Entailment Challenges (RTE). The measure shows to be good at discriminate over the "true" class. Therefore we develop a meta-classifier using a symmetric measure and a non-symmetric measure as base classifiers. So, our meta-classifier has a competitive performance.

1 Introduction

One of the biggest challenges in Natural Language Processing (NLP) is to provide a computer with the linguistic knowledge necessary to successfully perform language-based tasks. For example, the query "What does Peugeot manufacture?" a Question Answering (QA) system must be able to recognize, or infer, and answer which may be expressed differently from the query. Thus from text "Chrétien visited Peugeot's newly renovated car factory" entails the hypothesized answer from "Peugeot manufactures cars". A fundamental phenomenon in NLP is the variability of a semantic expression, which the same meaning could be expressed or infer from different text.

A task underlying this phenomenon is the ability to Recognize Textual Entailment. This task is defined as a directional relationship between pair of text expressions, denoted by T -the entailing "Text" and H -the entailed "Hypothesis". We say that T entails H if the meaning of H can be inferred from the meaning of T as could typically de interpreted by people [2].

Moreover, many NLP tasks have strong links to entailment: in Summarization (SUM), a summary should be entailed by the text; Paraphrases (PP) can be seen as mutual entailment between a text T and a hypothesis H; in Information Extraction (IE), the extracted information should also be entailed by the text; in QA the answer obtained for one question after the Information Extraction (IR) process must be entailed by the supporting snippet of text.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 372–381, 2010.

To address this task, different methods have been proposed, with various degrees of success. The classification of methods depends on the level of representation of the T-H pair. Therefore the common criteria for entailment recognition were similarity between T and H, or the coverage of H by T in lexical representation methods and lexical syntactic representation methods, and the ability to infer H from T, in the logical representation approach. Zanzotto et al also measured the similarity between different T-H pairs, crosspair similarity. Some works [6] tried to detect non-entailment, by looking for various kinds of mismatch between the text and the hypothesis.

In this paper we propose a new cause-effect non-symmetric measure for entailment recognition based on the causal relation between the text and the hypothesis. The causal relation is measure by using the relative frequencies of words in a cause-effect set. These sets are extracted from a corpus by searching sentences containing the discourse marker "because". Finally, we applied our method on a meta-classifier.

The paper is structured as follows. An overview of the related work in Section 2, Section 3 describes the proposed measure. Section 4 we shown experiments, and a comparison with previous results. Finally the conclusions are presented in Section 5.

2 Related Work

The RTE approaches can be classified depending in which textual entailment phenomena address or the type of representation (*levels of language*) of the T-H pair.

Thus each type of representation has operations in order to establish the entailment decision (e.g., word matching in the lexical level, tree edit distance in the syntactic level). The principal operations are similarity measures between T-H pair representations. But many of the similarity measures are symmetric. So a symmetric measure can not capture some of the aspects in the T→H relation. Because of if we altered the entailment relation (i.e., H→T) a symmetric function will give us the same score. Therefore methods like [9] propose a non-symmetric similarity measure, used in RTE-1 Challenge.

Glickman [3] uses as definition: T entails H iff $P(H|T) > P(H)$. The probabilities are calculated on the base of Web. The accuracy of the system is the best for RTE-1 (56%).

Another non-symmetric method is that of Kouylekov [7], who uses the definition: T entails H if and only if there exists a sequence of transformations applied to T such that H is obtained with a total cost below of a certain threshold. The following transformations are allowed: Insertion: insert a node from the dependency tree of H into the dependency tree of T; Deletion: delete a node from the dependency tree of T; Substitution: change a node in the T into a node of H. Each transformation has a cost and the cost of edit distance between T and H, $ed(T, H)$ is the sum of costs of all applied transformations. The entailment score of a given pair is calculated as

$$\text{score}(T,H) = \text{ed}(T,H),$$

where $ed(\cdot,H)$ is the cost of inserting the entire tree H. If this score is bigger than a learned threshold, the relation T →H holds. The accuracy of method is of 0.56.

In [9] an even "more non-symmetric" is proposed: when the edit distance (which is a Levenshtein modified distance) fulls the relation:

$$\text{ed}(T,H) < \text{ed}(H,T),$$

Then the relation T→H holds.

Other teams use a definition which in terms of representation of knowledge as feature structures could be formulated as: T entails H iff H subsumes T [9]. Even the method used in [2] is a non-symmetric one, as the definition used is: T entails H iff H is not informative in respect to T.

A method of establishing the entailment relation could be obtained using a non-symmetric measure of similarity between two texts presented by Corley and Mihalcea [1], the authors define the similarity between the texts T_i and T_j with respect to T_i as:

$$sim(T_i, T_j)_{Ti} = \frac{\sum_{pos}\left(\sum_{wk \in ws^{ti}_{pos}} (\max Sim(w_k) \times idf(w_k))\right)}{\sum_{pos}\sum_{wk \in ws^{ti}_{pos}} idf(w_k)}$$

Here the sets of open-class words (nouns, verbs, adjective and adverbs) in each text segment are denoted by WST_i PoS (PoS: Part of Speech) and WST_j PoS. For a word w_k with a given PoS in T_i, the highest similarity of the words with the same pos in the other text T_j is denoted by maxSim(w_k).

Starting with this text-to-text similarity metric, we derive a textual entailment recognition system by applying the lexical refutation theory presented above. As the hypothesis H is less informative than the text T, for a TRUE pair the following relation will take place:

$$\text{sim}(T,H) \times T < \text{sim}(T,H) \times H$$

This relation can be proven using the lexical refutation [9]. A draft is the following: to prove T→H it is necessary to prove that the set of formulas {T; negH} is lexical contradictory (they denote also by T and negH the sets of disjunctive clauses of T and negH).

3 Proposed Methods

A causal relation refers to the relation between a cause and its effect or between regularly correlated events. One type of coherence relation we used is cause-effect, illustrated above. For example: (1) states the cause for the effect given in (2).

1. *There was bad weather at the airport*
2. *and so our flight got delayed.*

The causal relation subsumes the cause and the explanation relations in Hobbs [3]. Hobbs's cause relation holds if a discourse segment stating a cause occurs before a discourse segment stating an effect; an explanation relation holds if a discourse segment stating an effect occurs before a discourse segment stating a cause. The causal

relation is encoded by adding a direction. In a graph, this can be represented by a directed arc going from cause to effect.

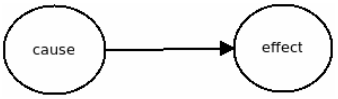

Fig. 1. Cause effect graph

Thus from Fig. 1 the causality is a directional relationship such as the relationship between a T-H pair. A non-symmetric similarity measure based on the count of co-occurrences of causal lexical pairs could be as follows: If a word *x* is a necessary cause of a word *y*, then the presence of *y* necessarily implies the presence of *x*.

3.1 Causal Non-symmetric Measure

The hypothesis behind our method is based on treat the T-H pair as a causal relation. Where the text T is a cause and the hypothesis H is its effect (i.e., T causes H).

The general scheme of the method is showed in Fig. 2:

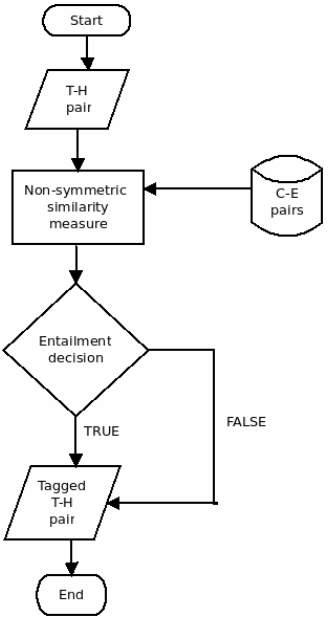

Fig. 2. General data flow of our system

In Fig. 2 we show the general data flow of the proposed method. The non-symmetric similarity measure is based on the count of co-occurrences of causal lexical pairs from a C-E pairs extracted from a corpus.

Algorithm 1. New non-symmetric similarity measure

```
For each word t_i in T
 For each word h_j in H
  ce_j=causal frequency(t_i,h_j)
  e_j=causal frequency(h_j)
 max_i = argmax(ce_j/e_j)
nonsymetric(T,H)= Σ max_i
```

As we se in the Algorithm 1 the first causal frequency function is the count of words t_i and h_i related by the cue phrase (For example, a sentence, h…because…t) in a corpus of C-E pairs and the second causal frequency function is the count of word h_i in the C-E pairs, which gives us a non-symmetric score. Because the co-occurrences of T causes H is not the same like H causes T.

To each T-H pair the system measures the causal relation between them and then decides if the pair is true or false given a certain entailment decision.

Algorithm 2. Entailment decision

```
if non-symmetric(T,H) > non-symetric(H,T) then TRUE
else FALSE
```

In Algorithm 2 we show that the entailment decision basically penalize a T—H pair when the H→T relation is stronger than the T→H relation. Therefore the hypothesis H is more probably an effect than the text T. Therefore it is more probable that the text T implies the hypothesis H.

3.2 Symmetric and Non-symmetric Meta-classifier

It has been observed for related systems that a combination of separately trained features in the machine learning component can lead to an overall improvement in system performance, in particular if features from a more informed component and shallow ones are combined.

One of the main problems when machine-learning classifiers are employed in practice is to determine whether classifications assigned to new instances are reliable. The meta-classifier approach is one of the simplest approaches to this problem. Given a base classifiers, the approach is to learn a meta-classifier that predicts the correctness of each instance classification of the base classifiers. The sources of the meta-training data are the training instances. The meta-label of an instance indicates reliable classification, if the instance is classified correctly by a base classifier; otherwise, the meta-label indicates unreliable classification. The meta-classifier plus the base classifiers form one combined classifier. The classification rule of the combined classifier is to

assign a class predicted by the base classifier to an instance if the meta-classifier decides that the classification is reliable.

Thus some questions on how to design a meta-classifier are:

- What type of base classifiers do we have to learn for meta-classifier, for what type of data?
- What is the role of the accuracy of the base classifiers in the whole scheme?
- How do we have to represent meta-data?
- How can we have to generate meta-data?

4 Experimental Setting

In this subsection we explain at detail some of the blocks in the Fig 2. First the preprocessing we used to represent the T-H pair and second the data used to create the C-E pairs.

The preprocessing we used in each T-H pair is as follows:

- Tokenize.
- Quit stop words.

Normally, an early step of processing is to divide the input text into units called tokens where each is either a *word* or something else like a number or a punctuation mark. This process is referred to as the treatment of punctuation varies.

The system has just stripped the punctuation out. We consider as word any object within the occurrence of a withespace. The withespace is the main clue used in English (RTE benchmark is in English). Finally the system quits any stops words from a stoplist. Common stop words are *the*, *from* and *could*. These words have important semantic functions in English, but they rarely contribute information if the criterion is a simple word-by-word match.

The data we used to collect the frequency of the causal lexical pairs came from sentences which contain the cue phrase *because.*). The sentences were striped in two parts: one corresponding to the cause and one corresponding to its effect to finally form the cause-effect pairs. The sentences were extracted from the Sketch Engine system over a big corpus (ukWAC from the Sketch Engine[1]). The Sketch Engine is a corpus query system which allows the user to view word sketches, thesaurally similar words, and 'sketch differences', as well as the more familiar Corpus Query Systems (CQS).

The answers to the questions of how to design a meta-classifier are as follows:

- We used symmetric and non-symmetric measures as base classifiers.
- We chose the best symmetric measure (we optimize accuracy).
- We represented the T-H pairs as a BoW.
- We used as meta-data the RTE Challenge test sets.

For the symmetric base classifier we tested between the cosine, word overlap, and the Bleu algorithm. Thus the cosine measure was the bet of all.

[1] `http://www.sketchengine.co.uk/`

5 Experimental Results

As we see in previous sections we varied the entailment decision in order to prove some differences between the uses of our non-symmetric measure. The experiment 1 was tested over the RTE-1 Challenge test set:

- Experiment 1: The system penalizes a pair if the H→T relation is greater than T→H relation.
- Experiment 2: The system determines the entailment decision based on a meta-classifier.

  The outline of the information displayed on each experiment is the next one:
- Contingency matrix.
- Evaluation matrix.
- Comparison with previous wok.
- Accuracy depending on task.

First, we present the method applied to the RTE-1. The contingency table, Table 3 show how many times the method misclassified the T-H pairs (i.e. *fp* and *tn*) and how many times the method its right. From this table we can obtain some measures to evaluate the entailment decision.

Table 3. RTE-1 contingency matrix

	true	false
true	257	245
false	143	155

Table 3 also shows that our approach tends to say true.

Table 4. RTE-1 evaluation measures

Accuracy	Precision	Recall	F-measure
0.51	0.51	0.64	0.57

From Table 4 this approach obtains a better recall than precision. Therefore the entailment decision got right the proportion of the target items that the system selected.

Table 5. RTE-1 comparison with previous results

Method	Accuracy
GLICKMAN	0.56
LEVENSHTEIN	0.53
C-E	0.51
BLEU	0.49

To compare our approach with previous works we use the accuracy measure (i.e. the most common measure in the RTE Challenge).The proposed measure is compared to non-symmetric measures. We compare out approach with:

- Bleu algorithm RTE baseline [8]
- Probabilistic measure [3]
- Levenshthein modified measure [9]

In Table 5 the results are show. Thus the best one is Glickman. Our measure is the last one compare to the non-symmetric measures. Our measure only outperforms the Bleu algorithm.

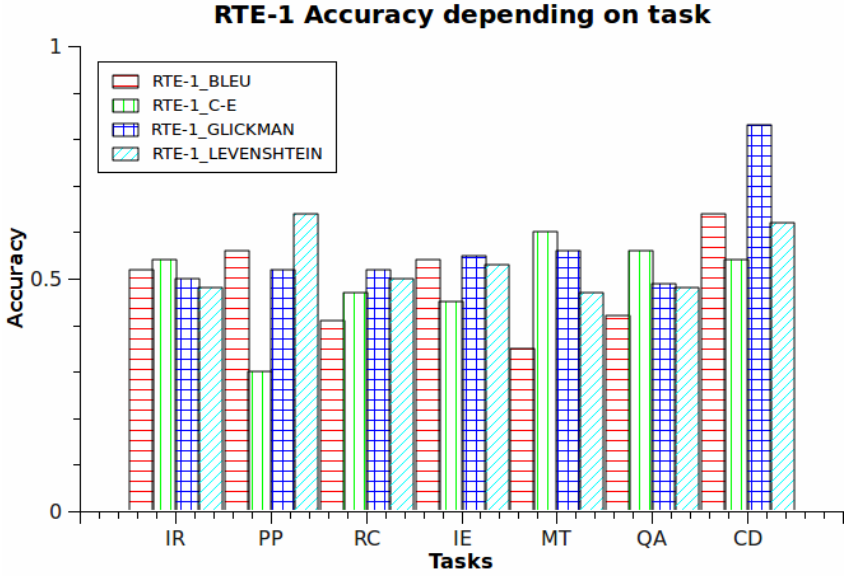

Fig. 3. RTE-1 comparison with previous results by tasks

The results of our approach were the lowest between the non-symmetric measures in general. So if we make a comparison depending on each task. We see that our measure outperforms the other non-symmetric measures in some of the tasks. These tasks are:

- QA.
- IR.
- MT.

The results of the meta-classifier over the RTE Challenge are: In the RTE-1 and RTE-2 the results did not achieve great differences against the Experiment 1. Thus in the RTE-3 the system achieve the best accuracy of all our experiments with 0.61.

In the RTE-3 we achieve the better results for our approach, comparing it to the other results in our research. Thus the results to the RTE-3 are competitive to other participants on the same Challenge.

The percentage of the coverage of the different base classifiers over the RTE-1 development data is as follows: Most of the T-H pairs could be resolved either by the symmetric and the non-symmetric measures (36.62%). Following the examples

resolved by the symmetric measure (29.38%) and the non-symmetric at last (14.12%). Finally the 18.88% of the instances could not be resolved by any measure.

Table 6. RTE-3 meta-classifier contengiency matrix

	true	false
true	264	163
false	146	227

Table 7. RTE-3 meta-classifier evaluation measure

Accuracy	Precision	Recall	F-measure
0.61	0.61	0.64	0.63

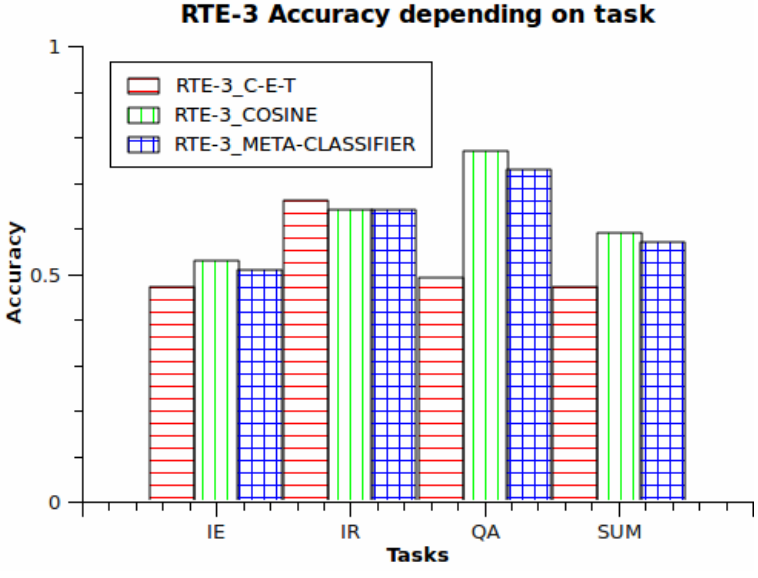

Fig. 4. RTE-3 meta-classifier comparison with base classifiers by tasks

6 Conclusion and Future Work

We proposed a non-symmetric similarity measure to the RTE task. Therefore our unsupervised method is no language dependent.

We have shown that our measure has a lower accuracy than the state of the art methods and outperforms the RTE baseline. These results are significant because they are based on a very simple algorithm that relies on co-occurrences of causal pairs.

We once more confirmed that the web could be used as a lexical resource for RTE (i.e. The Sketch Engine developers have built their corpora from the Web). Also our meta-classifier has a competitive accuracy of 0.61; the average accuracy for the RTE-3 is of 0.61.

In our future work we will explore the use of different meta-features for the meta-classifier, as well as linguistically-motivated meta-features (such as a syntactic unit) and evaluate our method against the RTE machine learning approaches.

References

1. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, pp. 13–18 (June 2005)
2. Dagan, I., Glickman, O.: Probabilistic textual entailment: Generic applied modeling of language variability. In: PASCAL workshop on Text Understanding (2004)
3. De Salvo Braz, R., Girju, R., Punyakanok, V., Frentiu, D.M.: An Inference Model for Word Sense Disambiguation. In: Proceedings of KEPT 2007, Knowledge Engineering Principles and Techniques, Workshop on Recognising Textual Entailment, vol. I (2007)
4. Glickman, O., Dagan, I., Koppel, M.: Web Based Probabilistic Textual Entailment. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (2005)
5. Hobbs, J.R.: Ontological promiscuity. In: Proceedings of the 23rd annual meeting on Association for Computational Linguistics (1985)
6. Inkpen, D., Kipp, D., Nastase, V.: Machine Learning Experiments for Textual Entailment. In: Proceedings of the Second Challenge Workshop Recognising Textual Entailment, Venice, Italy (2006)
7. Kouylekov, M., Magnini, B.: Tree Edit Distance for Recognizing Textual Entailment: Estimating the Cost of Insertion. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy (2006)
8. Pérez, D., Alfonseca, E.: Application of the Bleu algorithm for recognising textual entailments. In: Proceedings of the First Challenge Workshop Recognising Textual Etailment, Southampton, U.K., April 11-13, pp. 9–12 (2005)
9. Tatar, D., Gabriela, S., Andreea-Diana, M., Rada, M.: Textual Entailment as a Directional Relation. Journal of Research and Practice in Information Technology (2009)