IKOMA at TAC2010: Textual Entailment System using Local-Novelty Detection

Kenji TATEISHI and Kai ISHIKAWA
Information and Media Processing Laboratories, NEC Corporation
{k-tateishi@bq, k-ishikawa@dq}.jp.nec.com

Abstract

This paper reports the Recognizing Textual Entailment (RTE) system that our '*IKOMA team*' developed at TAC 2010. We implemented a new method that identifies entailment sentences using local-novelty detection. This method determines whether a Hypothesis (H) is local-novel. '*H is local-novel*' means that H is the new information first appeared in T_H, and T_H denotes the text (T) that entails H. If H is local-novel, the T that was published before T_H can be recognized as no entailment. The experimental results show that this method detects T that does not entail H correctly without setting similarity threshold to be high, and raises precision as minimizing the decline of recall.

1. Introduction

A Recognizing Textual Entailment (RTE) task determines whether a given Text (T) entails a given Hypothesis (H). Based on task guideline [1], T entails H if, typically, a human reading T would infer that H is most likely true. For example, if "President Barack Obama visited Japan" is given as H and "President Obama met Japanese Prime Minister in Tokyo" is given as T, the RTE system should answer *ENTAILMENT=YES*, T entails H.

Many RTE systems use a similarity-based matching method [2,3] that is based on the assumption that the possibility of entailment can be substituted for the value of the similarity score between H and T. For example, the system in [2] concludes *ENTAILMENT=YES* if the percentage of common dependency-pairs (triplets) between H and T (similarity score) exceeds a given similarity threshold. Also, this system expands this basic method using language resources and semantic rules.

The problem with this similarity-based approach is its low performance. Generally, the similarity threshold has to be set high to prevent false-positives (to maintain a high precision rate). However, it increases false-negatives (decline of recall). Since every sentence can be expressed in various ways, the similarity score is not always high even in the case of *ENTAILMENT=YES*.

In this paper, we propose a new method that identifies entailment sentences using local-novelty detection. Our proposed method first determines whether H is local-novel. *'H is local-novel*' means that H is the new information first appeared in T_H, and T_H denotes the T that entails H. If H is local-novel, the T that was published before T_H never entails H. Therefore, the T can be recognized as *ENTAILMENT=NO*. It works as pre-processing of the similarity-based matching method. Our proposed method correctly detects T that does not entail H without setting a high similarity threshold, and raises the precision as minimizing the decline of recall.

Ex. 1)
H: Peter Jennings died at the age of 67.
T_H: ABC News anchor Peter Jennings died of lung cancer at the age of 67 on Sunday, his company said.

→ Local-Novel

Ex. 2)
H: Gerry Adams is the leader of Sinn Fein.
T_H: Gerry Adams, the leader of the Sinn Fein political wing of the Irish Republican Army (IRA), confirmed Thursday that the Roman Catholic paramilitary group was on the verge of making a significant statement.

→ Not Local-Novel

Figure 1: Examples of Local-Novel Detection

it in Section 3. In Section 4, we explain related work and conclude in Section 5.

2. Entailment Judgment using Local-Novelty Detection(LND)

Our proposed method first determines whether H is local-novel. '*H is local-novel'* means that H is the new information first appeared in T_H, and T_H denotes the T that entails H. If H is local-novel, the T that was published before T_H never entails H. Therefore, the T can be recognized as *ENTAILMENT=NO*. The proposed method correctly detects T that does not entail H, and raises precision as minimizing the decline of recall.

The proposed method can be applied to a document set that satisfies these two conditions: (a) each document has a label of its publishing date. (b) T_H, the T that entails H, is given beforehand. As we explain on Section 3.1, the test set on RTE-6 fulfils both.

The proposed method processes in two steps. In Step 1, the system determines whether H is local-novel, and in Step 2, the system identifies T that entails H using the Step 1 results.

2.1 Step 1: Local-Novelty Detection

'H is local-novel' means that H is the new information first appeared in T_H, and T_H denotes the T that entails H. Ex.1 in Fig. 1 is an example where H is likely to be local-novel because the information provider of H (*his company said*) is specified on T_H, and newswire articles usually add providers at the point at which the information first appears. On the other hand, Ex.2 in Fig.1 is an example where H is unlikely to be local-novel. This is because H is located in a parenthetical clause on T_H, and H is treated as unimportant. H is likely to be a general fact or published information.

Local-novel detection module processes with decision-rules. It uses the six decision-rules in Tab. 1, and applies them from **[R1]** to **[R6]**. These rules target newswire articles. If neither rule is satisfied, the result is unknown. The following provides the details.

[R1] If H is located in parenthetical clause or subordinate clauses on T_H, H is not local-novel.

The H located in parenthetical or subordinate clauses in T_H is treated as unimportant in the document, and H is likely to be a general fact or previously published information. The module

splits T_H into blocks by commas. Then it judges whether a block is a subordinate clause if the part-of-speech of the block's head word is a conjunction or a preposition. Also, if the previous block is not judged to be a subordinate clause and the part-of-speech of the block's head word is a noun or an article, it judges the block to be a parenthetical clause. Then if the block in which the verb in H included is located in the parenthetic or subordinate clause, it determines that H is not local-novel.

[R2] If the sentence type of H is "A is B.", H is not local-novel.

An "A is B." type of sentence is likely to be a general fact. If H does not include a verb other than a form of 'to be', the module determines that H is not local-novel.

[R3] If the information provider of H is specified in T_H, H is local-novel.

Newswire articles usually add providers at the point at which the information first appears. If T_H includes "announced", "reported", "said", "according to", or "told", the module determines that H is local-novel.

[R4] If T_H includes a time expression that indicates the long past, H is not local-novel.

A fact occurred in the long past is probably previously published information. If the block where the verb in H is included contains a month expression (ex. January, February) or year expression (ex. 2009), it determines that H is not local-novel.

[R5] If the document title of T_H entails H, H is local-novel.

The document title (ex. newswire article headline) is likely to show new information. The module uses TBM in Section 3.3 to judge whether the document title entails H.

[R6] T_H just includes H, H is local-novel.

Newswire articles typically provide new information to public. Therefore, if T_H just simply includes H, H is likely to have local-novelty. The module determines that H is local-novel if the cosine similarity between H and T_H is equal or exceeds 0.9. The term weight function on cosines similarity follows w(t) of TBM in Section 3.3.

Table 1: Six Decision-Rules for Local-Novelty Detection

ID	Priority	IF	THEN
R1	1	H is located in parenthetical clause or subordinate clause on T_H	NOT Local-Novel
R2	2	Sentence type of H is "A is B."	NOT Local-Novel
R3	3	Information provider of H is specified on T_H	Local-Novel
R4	4	T_H includes a time expression that indicates the long past	NOT Local-Novel
R5	5	Document title of T_H entails H	Local-Novel
R6	6	T_H just includes H	Local-Novel

```
def EntailmentJudgment(H, T_H, T, baseThreshold)
        th = baseThreshold
        if detectLocalNovelty(H, T_H) == true
                if getDate(T) < getDate(T_H)
                        th += x
                end
        end
        if calcSimilarity(H, T) >= th
                return true
        else
                return false
        end
end
```

Figure 2: Function of Entailment Judgment

2.2 Step 2: Entailment Judgment

In Step 2, the system identifies T that entails H using the Step 1 result. If the system determines that H is local-novel in Step 1, and if T was published before T_H, then it adds *x* point to the initial similarity threshold. Note that if the infinite value is set as *x*, T never entails H, regardless of the similarity score between H and T.

Table 2 shows the pseudocode of the the entailment judgment, where *baseThreshold* is a constant value that denotes the similarity threshold, *detectLocalNovelty* is a function that determines whether H is local-novel, *getDate* is a function that obtains the T's or H's publishing date, and *calcSimilarity* is a function that calculates the similarity score between H and T.

3. Experiment

3.1 Data Set

The main application target on RTE-6 is the Update Summarization Task (UST) [1], which summarizes the documents of *Cluster B* under the condition that the user has already read the documents of *Cluster A*. RTE can be used for detecting candidate sentences or phrases for summarization, when the sentences or phrases of Cluster B are treated as Hs and the sentences of Cluster A are treated as Ts.

We used the test set given to the participants of RTE-6 [1]. Each document set of the test set, *Clusters A* or *B*, is composed of 10 topics x 10 documents = 100 documents. H is a sentence or a phrase included in *Cluster B*. Each H has 100 Ts at most, all of which were obtained by retrieving the top 100 sentences in *Cluster A* by giving the H as the query. There are 243 Hs and 19972 Ts. Each participant is given the development set in addition to the test set, which has the same data format and has almost the same quantity as the test set.

The test set on RTE-6 fulfills the two conditions mentioned in Section 2.1: (a) each document/newswire article has a publishing date, and (b) T_H is given, and is the sentence in

Cluster B that includes H. Due to the nature of UST, the documents in *Cluster A* are always published before those in *Cluster B*. This means that T is always published before T_H.

3.2 Evaluation Criterion

RTE-6 has two tasks, a Main Task and a Novelty Detection Sub-Task, both of which evaluate its performance by F-measure. The main-task is an entailment judgment task whose input is an H and whose output is Ts that entails the H. The correct data is defined as Ts that entail an H. The sub-task is a novelty-detection task whose input is an H and whose output is whether the H is NOT written in Cluster A. The correct data is defined as Hs that have no T that entails them.

Here, F-mesure is the harmonic average of precision and recall. The macro average is defined as the average of the F -values by each topic on the test set, and the micro average indicates the F-measure straightforwardly obtained from the whole test set.

3.3 Baseline

First, we selected a baseline by comparing the following two methods.

[TBM] Term-Based Matching method

Sim(H,T), the similarity score between H and T, is defined as the following formula.

$$\mathrm{Sim}(H,T) = \frac{\sum_{t \subset H \cap T} \mathrm{w}(t)}{\sum_{t \subset H} \mathrm{w}(t)}$$

$$\mathrm{w}(t) = \log_2 \frac{|T|}{\mathrm{textfreq}(t)}$$

Here, $t \subset$H denotes the the terms included in H, $t \subset$H$\cap$T indicates the terms that commonly appear in H and T, |T| is the total number of Ts, w(t) is the weight of t, and textfreq(t) is the number of Ts that include t. “a” and “the” is used as stopwords.

[PBM] Pair-Based Matching method.

Sim(H,T), the similarity score between H and T, is defined as the following formula.

$$\mathrm{Sim}(H,T) = \frac{\sum_{dp \subset H \cap T} \mathrm{w}(dp)}{\sum_{dp \subset H} \mathrm{w}(dp)}$$

$$\mathrm{w}(dp) = \max\{\mathrm{w}(t_1), \mathrm{w}(t_2)\}$$

Here, $dp \subset$ H denotes the dependency-pairs included in H, $dp \subset$ H$\cap$T indicates the dependency-pairs that commonly appear in H and T, w(dp) is the weight of dp, and t_1, t_2 are the two terms that composes dp. “a” and “the” are used as stopwords. To create dependency pairs, we used MINIPAR[5] as a syntactic analysis tool.

Table 2 compares TBM and PBM. In this experiment, we determined similarity threshold *th* from the development set. We sought the best similarity threshold by changing it at 0.05 intervals

and selected *th*=0.40 in TBM and *th*=0.15 in PBM. From this result, we chose TBM as the baseline.

Table 2: Comparison between TBM and PBM

F-measure	TBM (*th*=0.40)	PBM (*th*=0.15)
Micro Ave.	**45.27**	30.29
Macro Ave.	**45.86**	31.91

3.4 Experimental Result

Table 3 and Table 4 compare TBM and our proposed method on the main- and the sub-tasks. Note that the proposed method set *x*=0.2, which is a constant that adds to the initial similarity threshold, and used TBM as *calcSimilarity*, which is a function that calculates the similarity between H and T. We determined the (initial) similarity threshold from the development set by changing it at 0.05 intervals. The best value was *th*=0.40 in the main-task, and *th*=0.45 in the sub-task.

Table 3: Comparison between TBM and Proposed Method on Main Task.

F-measure	TBM (*th*=0.40)	TBM+LND (*th*=0.40)
Micro Ave.	45.27	**45.41**
Precision	38.72	40.92
Recall	54.50	51.01
Macro Ave.	45.86	**46.01**
Precision	29.40	41.55
Recall	54.87	51.54

Table 4: Comparison between TBM and Proposed Method on Novelty Detection Sub-Task.

F-measure	TBM (*th*=0.45)	TBM+LND (*th*=0.45)
Micro Ave.	80.61	**82.13**
Precision	**82.29**	79.44
Recall	79.00	**85.00**
Macro Ave.	81.17	**82.53**
Precision	**84.50**	80.13
Recall	78.10	**85.08**

3.5 Discussion

The F-mesure of the proposed method exceeded that of the baseline in both the main- and sub-tasks. In the sub-task, the decline of precision was 2.85-4.37 points and the raise of recall was 6.00-6.98 points. This result indicates the proposed method can correctly identify T that does not entail H without giving a high similarity threshold, and raises precision as preventing the decline of recall in the entailment judgment task.

Our proposed method was more effective in the sub-task compared with baseline than in the main-task .This can be considered that regarding the H that the proposed method determines is

local-novel, there are few number of Ts that the baseline identifies entails the H. Since the correct data of the main-task is Ts that entails an H, there is little room to raise the F-measure with the proposed method in the main-task. On the other hand, since the correct data of sub-task is Hs that has no T that entails them, there is room to raise the F-measure with the proposed method.

Table 5 shows the performance of the local-novelty detection module. For constructing the experimental data, we treated H with no T that entails it as local-novel on the development set, and H with more than one Ts that entail it as not local-novel. Note that since H is not always local-novel even if H has no T that entails it, this result is roughly estimated. The performance of the local-novelty detection is 73% and exceeded the baseline's 42%- in the simple method, all Hs are determined as local-novel. The decline of precision in the sub-task was caused by false-positive, which are errors of local-novelty detection. We need to refine the decision-rules in Section 2.21 (ex. adding decision-rules or changing their priority in Section 2.1 in the future work.

Table 5 Precision of Proposed Method on Local Novelty Detection Module

	LND	Baseline
Local-Novel	0.73(35/48)	0.42(88/211)
NOT Local-Novel	0.83(83/100)	0.58(123/211)

3.6 Effect of Other Functions

Table 6 compares TBM and the following methods. We selected similarity threshold *th*=0.40.

[TBM+Acronym] TBM + Acronym Extraction

The system first runs the NER tool on all sentences in test set. We used Stanford NER [6]. Then, it creates acronyms by selecting the initial alphabet of each word if labeled "ORGNIZATION" and if composed of more than three words (ex. Irish Republican Army → IRA), and by selecting a second word (family name) if labeled "PERSON" and if composed of two words (Ex. Casey Sheehan → Sheehan). Then, it applies TBM after it unifies the organization and person expressions using these acronyms.

[TBM+TopicWeight] TBM + Term Weight Modification

The system changes w(t) on TBM to w'(t) to give a large weight to terms that appear in a few topics.

$$\mathrm{w}'(t) = \log_2 \frac{|T|}{\mathrm{textfreq}(t) \times \left(1 + \frac{\mathrm{topicfreq}(t)}{\mathrm{topicnum}}\right)}$$

Here, topicfreq(t) denotes the number of topics in which t appears in more than one document, and topicnum is athetotal number of topics (=10).

[TBM+WordNet] TBM+WordNet

The system first searches for synonyms of each word (noun, verb, adjective and adverb)

registered on WordNet [7] by finding the primal synset (word group with highest priority) of the word. Then it unifies the expressions in the test set with the synonyms, and applies TBM.

Table 6 Comparison between TBM and Other Functions.

F-measure	TBM	TBM+Acronym	TBM+TopicWeight	TBM+WordNet
Micro Ave.	45.27	45.20	45.39	45.48
Macro Ave.	45.86	45.44	46.30	46.01

3.7 Formal-Run Result

We submitted three runs: IKOMA1, IKOMA2 and IKOMA3. All of them are combination of baseline method, local-novelty detection, and other functions.

Table 7 Formal-Run Result (Main-Task)

Run	Method	Similarity Threshold	F-measure (Micro Ave.)	F-measure (Macro Ave.)
IKOMA1	TBM+LND+Acronym	0.40	44.81	45.11
IKOMA2	TBM+LND+Acronym+TopicWeight	0.45	44.78	44.91
IKOMA3	TBM+LND+Acronym+TopicWeight	0.45	44.59	45.40

Table 8 Formal-Run Result (Sub-Task)

Run	Method	Similarity Threshold	F-measure (Micro Ave.)
IKOMA1	TBM+LND	0.40	77.25
IKOMA2	TBM+LND	0.45	82.13
IKOMA3	TBM+LND	0.50	80.93

Table 9 Result of Ablation Test (Main-Task)

Run	Ablated Module	F-measure	Contribution
IKOMA2	-	44.78	-
IKOMA2_abl-1	TopicWeight	44.81	-0.03
IKOMA2_abl-2	LND	44.71	+0.07
IKOMA2_abl-3	Acronym	45.54	-0.76

4. Related Work

Through the RTE Challenge much research work on RTE has been done. There are two approaches: a similarity-based matching method [2,3] and a machine-learning-based matching method [4] that classifies pairs of Hs and Ts into two categories. In Section 3, we compared the proposed method with a similarity-based matching method and showed the effectiveness. The proposed method can also be combined with the latter by using it as pre-processing.

Since RTE-6 changed its task description from the previous RTE challenges to resemble the IR task, the method showed effectiveness up to RTE-5 is not always effective on RTE-6. In fact, PBM, which is employed by many top-ranked terms, was inferior to TBM- more simple approach.

Sentences can be expressed in various ways, so the H's syntactic structures do not remain in H. Similarly, using WordNet or acronym extraction was not effective so much either.

5. Conclusion

This paper reported the RTE system using local-novelty detection that our '*IKOMA* team' developed in TAC 2010. The experimental results showed that our proposed method detected T that does not entail H correctly, and raised precision as minimizing the decline of recall.

References

[1] 6th TEXTUAL ENTAILMENT CHALLENGE@TAC 2010 MAIN TASK and NOVELTY DETECTION SUBTASK TASK Guideline, http://www.nist.gov/tac/2010/RTE/RTE6_Main_NoveltyDetection_Task_Guidelines.pdf

[2] Adrian Iftene, TEXTUAL ENTAILMENT, http://profs.info.uaic.ro/~adiftene/thesisAI.pdf

[3] Rui Wang, Guenter Neumann, An Divide-and Conquer Strategy for Recognizing Textual Entailment, Proceedings of TAC 2008, http://www.nist.gov/tac/publications/2008/participant.papers/DFKI.proceedings.pdf

[4] Jeremy Bensley and Andrew Hickl, Workshop: Application of LCC's GROUNDHOG System for RTE-4, Proceedings of TAC 2008, http://www.nist.gov/tac/publications/2008/participant.papers/lcc.proceedings.pdf

[5] MINIPAR, http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

[6] Stanford Named Entity Recognizer (NER), http://nlp.stanford.edu/software/CRF-NER.shtml

[7] WordNet, http://wordnet.princeton.edu/