Polibits
ISSN: 1870-9044
polibits@nlp.cic.ipn.mx
Instituto Politécnico Nacional
México

Pakray, Partha; Poria, Soujanya; Bandyopadhyay, Sivaji; Gelbukh, Alexander
Semantic Textual Entailment Recognition using UNL
Polibits, vol. 43, 2011
Instituto Politécnico Nacional
Distrito Federal, México

Semantic Textual Entailment Recognition using UNL

Partha Pakray, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh

Abstract—**A two-way textual entailment (TE) recognition system that uses semantic features has been described in this paper. We have used the Universal Networking Language (UNL) to identify the semantic features. UNL has all the components of a natural language. The development of a UNL based textual entailment system that compares the UNL relations in both the text and the hypothesis has been reported. The semantic TE system has been developed using the RTE-3 test annotated set as a development set (includes 800 text-hypothesis pairs). Evaluation scores obtained on the RTE-4 test set (includes 1000 text-hypothesis pairs) show 55.89% precision and 65.40% recall for YES decisions and 66.50% precision and 55.20% recall for NO decisions and overall 60.3% precision and 60.3% recall.**

Index Terms—**Textual Entailment, Universal Networking Language (UNL), RTE-3 Test Annotated Data, RTE-4 Test Data**

I. Introduction

RECOGNIZING Textual Entailment is one of the recent challenges of Natural Language Processing. Textual Entailment is defined as a directional relationship between pairs of text expressions, denoted by the entailing “Text” (T) and the entailed “Hypothesis” (H). T entails H if the meaning of H can be inferred from the meaning of T.

Textual Entailment has many applications in Natural Language Processing tasks: in Summarization (SUM), a summary should be entailed by the text; Paraphrases (PP) can be seen as mutual entailment between a T and a H; in Information Extraction (IE), the extracted information should also be entailed by the text; in Question Answering (QA) the answer obtained for one question after the Information Retrieval (IR) process must be entailed by the supporting snippet of text.

There were three Recognizing Textual Entailment competitions RTE-1 in 2005 [4], RTE-2 [1] in 2006 and RTE-3 [6] in 2007 which were organized by PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning)—European Commission’s IST-funded Network of Excellence for Multimodal Interfaces. In 2008, the fourth edition (RTE-4) of the challenge was organized by NIST (National Institute of Standards and Technology) in Text Analysis Conference (TAC). In every new competition several new features of RTE were introduced. The RTE-5 challenge in 2009 includes a separate search pilot along with the main task.

The first PASCAL Recognizing Textual Entailment Challenge (RTE-1) [4], introduced the first benchmark for the entailment recognition task. The RTE-1 dataset consists of manually collected text fragment pairs, termed text *t* (1-2 sentences) and hypothesis *h* (one sentence). The systems were required to judge for each pair whether *t* entails *h*. The pairs represented success and failure settings of inferences in various application types (termed “tasks”). In RTE-1 the various techniques used by the participating systems were word overlap, WordNet, statistical lexical relation, world knowledge, syntactic matching and logical inference.

After the success of RTE-1, the main goal of the RTE-2, held in 2006 [1], was to support the continuity of research on textual entailment. The RTE-2 data set was created with the main focus of providing more “realistic” text-hypothesis pair. As in the RTE-1, the main task was to judge whether a hypothesis H is entailed by a text. The texts in the datasets were of 1-2 sentences, while the hypotheses were one sentence long. Again, the examples were drawn to represent different levels of entailment reasoning: lexical, syntactic, morphological and logical. The main task in the RTE-2 challenge was classification—entailment judgment for each pair in the test set that represented either entailment or no entailment. The evaluation criterion for this task was accuracy—the percentage of pairs correctly judged. A secondary task was created to rank the pairs based on their entailment confidence. A perfect ranking would place all the positive pairs (for which the entailment holds) before all the negative pairs. This task was evaluated using the average precision measure [8], which is a common evaluation measure for ranking in information retrieval. In RTE-2 the techniques used by the various participating systems are Lexical Relation/ database, n-gram/ subsequence overlap, syntactic matching/ Alignment, Semantic Role labeling / FrameNet / PropBank, Logical Inference, Corpus/web-based statistics, machine learning (ML) Classification, Paraphrase and Templates, Background Knowledge and acquisition of entailment corpus.

The RTE-3 data set consisted of 1600 text-hypothesis pairs, equally divided into a development set and a test set. The same four applications from RTE-2—namely IE, IR, QA and

Manuscript received November 2, 2010. Manuscript accepted for publication January 12, 2011. This work was supported in part by the Government of India and Government of Mexico (joint DST-CONACYT project) and Government of Mexico (CONACYT 50206-H, SIP-IPN 20113295, as well as SNI and CONACYT Sabbatical program as to the fourth author).

P. Pakray, S. Poria, and S. Bandyopadhyay are with the Computer Science and Engineering Department, Jadavpur University, Kolkata, India (e-mail: parthapakray@gmail.com, soujanya.poria@gmail.com, sbandyopadhyay@cse.jdvu.ac.in).

A. Gelbukh is with the Faculty of Law, Waseda University, Tokyo, Japan, on Sabbatical leave from the Center for Computing Research, National Polytechnic Institute, Mexico City, Mexico (e-mail: gelbukh@gelbukh.com).

SUM—were considered as settings or contexts for the pair's generation. 200 pairs were selected for each application in each data set. Each pair was annotated with its related task (IE/IR/QA/SUM) and entailment judgment (YES/NO). In addition, an optional pilot task, called "Extending the Evaluation of Inferences from Texts" was set up by the NIST, in order to explore two other sub-tasks closely related to textual entailment: differentiating unknown entailment from identified contradictions and providing justifications for system decisions. In the first sub-task, the idea was to drive systems to make more precise informational distinctions, taking a three-way decision between "YES", "NO" and "UNKNOWN", so that a hypothesis being unknown on the basis of a text would be distinguished from a hypothesis being shown false/contradicted by a text.

In RTE-4 [5], no development set was provided, as the pairs proposed were very similar to the ones contained in RTE-3 development and test sets, which could therefore be used to train the systems. Four applications—namely IE, IR, QA and SUM—were considered as settings or contexts for the pair generation. The length of the H's was the same as in the past data sets (RTE-3); however, the T's were generally longer. A major difference with respect to RTE-3 was that the RTE-4 data set consisted of 1000 T-H pairs, instead of 800. In RTE-4, the challenges were classified as two-way task and three-way task. The two-way RTE task was to decide whether:

1) T entails H—in which case the pair will be marked as ENTAILMENT;
2) T does not entail H—in which case the pair will be marked as NO ENTAILMENT.

The three-way RTE task was to decide whether:

3) T entails H—in which case the pair was marked as ENTAILMENT,
4) T contradicts H—in which case the pair was marked as CONTRADICTION,
5) The truth of H could not be determined on the basis of T—in which case the pair was marked as UNKNOWN.

In every new competition several new features of RTE were introduced. The TAC RTE-5 [2] challenge in 2009 includes a separate search pilot along with the main task. The TAC RTE-6 challenge[1], in 2010, includes the Main Task and Novelty Detection Task along with RTE-6 KBP Validation Pilot Task. The RTE-6 does not include the traditional RTE Main Task, which was carried out in the first five RTE challenges, i.e., there was no task to make entailment judgments over isolated T-H pairs drawn from multiple applications. In 2010, Parser Training and Evaluation using Textual Entailment [9] was organized by SemEval-2. We have developed our own RTE system and have participated in TAC RTE-5 and Parser Training and Evaluation using Textual Entailment as part of SemEval-2 and also in TAC RTE-6.

In the present paper, a 2-way semantic textual entailment recognition system has been described that has been trained on the 2-way RTE-3 test gold set and then tested on the RTE-4 test set. UNL Expressions are described in Section 2. Section 3 describes semantic based RTE system architecture. The experiment carried out on the development and test data sets are described in Section 4 along with the results. The conclusions are drawn in Section 5.

1 http://www.nist.gov/tac/2010/RTE/index.html

II. UNL EXPRESSIONS

Universal Networking Language (UNL) is an artificial language that can be used as a pivot language in machine translation systems or as a knowledge representation language in information retrieval applications. The UNL [3, 7] expresses information or knowledge in the form of semantic network with hyper-node. UNL semantic network is made up of a set of binary relations, each binary relation is composed of a relation and two Universal Words (UWs) that hold the relation. A binary relation of UNL is expressed in the format shown in Table I.

TABLE I
UNL RELATION

<relation> (<uw1>, <uw2>)

In <relation>, one of the relations defined in the UNL Specifications is described. In <uw1> and <uw2>, the two UWs that hold the relation given at <relation> are described.

All binary relations that compose a UNL expression have directions, and the semantic network of a UNL expression is a directed hyper-graph.

A. UNL expression hyper-graph

Each UNL expression is a semantic hyper-network. That is, each node of the graph, <uw1> and <uw2> of a binary relation, can be replaced with a semantic network. Such a node consists of a semantic network of a UNL expression and is called a "scope". A scope can be connected with other UWs or scopes. Each UNL expression in a scope is distinguished from others by assigning an ID to the <relations> of the set of binary relations that belong to the scope.

The general description format of binary relations for a hyper-node of UNL expression is in Table II, where:

– <scope-id> is the ID for distinguishing a scope. <scope-id> is not necessary to be specified when a binary relation does not belong to any scope.
– <node1> and <node2> can be a UW or a <scope node>.
– A <scope node> is given in the format ": <scope-id>".

TABLE II
UNL EXPRESSION

<relation>:<scope-id> (<node1>, <node2>)

An example UNL expression for hypothesis is given in Table III.

The EnConverter and DeConverter are the core software in the UNL System. The EnConverter converts natural language sentences into UNL Expressions. The DeConverter converts

TABLE III
UNL RELATION

```
{org:en}
UN peacekeepers abuse children.
{/org}
{unl}
mod(peacekeeper(icl>defender>thing).@pl,un(icl>world_organization>
    thing,equ>united_nations))
agt(abuse(icl>treat>do,equ>mistreat,agt>person,obj>living_thing).@entry.
    @present,peacekeeper(icl>defender>thing).@pl)
obj(abuse(icl>treat>do,equ>mistreat,agt>person,obj>living_thing).@entry.
    @present,child(icl>juvenile>thing).@pl)
{/unl}
```

UNL Expressions to natural language sentences. Both the EnConverter and DeConverter perform their functions based on a set of grammar rules and a word dictionary of a target language.

B. UNL Relations

Some of the UNL Relations are shown in Table IV. We used the Expanded Rules in Table VIII. These expanded rules, based on the present UNL Expression, have been developed from the RTE-3 test annotated corpus. Then these rules are applied on RTE-4 test set. Currently the system has 35 expanded rules.

TABLE IV
UNL RELATION DESCRIPTION

Relations Name	Details
agt (agent)	defines a thing that initiates an action.
mod (modification)	defines a thing that restricts a focused thing.
nam (name)	defines a name of a thing.
plc (place)	defines a place where an event occurs, or a state that is true, or a thing that exists.
plt (final place)	defines a place where an event ends or a state that is false.
tim (time)	defines the time an event occurs or a state that is true.
tmf (initial time)	defines the time an event starts or a state that is true.
tmt (final time)	defines a time an event ends or a state that is false.
to (destination)	defines a final state of a thing or a final thing (destination) associated with the focused thing.
src (source: initial state)	defines the initial state of an object or thing initially associated with the object of an event.
obj(affected thing))	defines a thing in focus that is directly affected by an event or state.

III. SYSTEM DESCRIPTION

In this section, we describe our semantic based textual entailment system. The system accepts pairs of text snippets (text and hypothesis) at the input and gives a value at the output: YES if the text entails the hypothesis and NO otherwise. The architecture of the proposed system is described in Fig. 1.

A. Pre-processing

The system accepts pairs of text snippets (text and hypothesis) at the input and gives the output: YES if the text entails the hypothesis and NO otherwise. An example text-hypothesis pair from the RTE-3 test annotated set which is used as a development set is shown in Table V.

TABLE V
RTE-3 TEST ANNOTATED SET

```
<pair id="12" entailment="YES" task="IE" length="short" >
<t>Judge Drew served as Justice until Kennon returned to claim his seat in
    1945.</t>
<h>Kennon served as Justice.</h>
</pair>
```

In the development set, the following expressions were replaced: “aren’t” with “are not”, “didn’t” with “did not”, “doesn’t” with “does not”, “won’t” with “will not”, “don’t” with “do not”, “hasn’t” with “has not”, “isn’t” with “is not”, “couldn’t” with “could not”, “ă” with “a”, "á" with “a”, "š" with “s”, "ž" with “z” and "ó" with “o”. These expressions are either abbreviations or include special characters for which the dependency parser gives erroneous results. It has also been observed that escape characters like ", …, ‘ and & are present in the text and the hypothesis parts and these were removed. All the above pre-processing methods were also applied on the RTE-4 test set.

B. UNL Enconverter Module

In this module, we convert the text and hypothesis pair into UNL expressions[2]. For example, the UNL expression for the hypothesis in Table V is shown in Table VI, and the UNL Graph for this hypothesis is shown in Fig. 2.

TABLE VI
UNL EXPRESSION FOR RTE-3 TEST ANNOTATED SET HYPOTHESIS

```
[S:00]
{org:en}
Kennon served as Justice
{/org}
{unl}
aoj(serve(icl>be,obj>uw,aoj>thing,ben>thing).@entry.@past,kennon)
obj(serve(icl>be,obj>uw,aoj>thing,ben>thing).@entry.@past,justice
    (icl>righteousness>thing,ant>injustice).@maiuscul)
{/unl}
[/S]
```

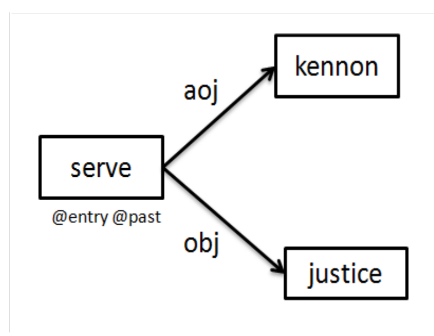

Fig. 2. UNL Hyper-graph.

In this case the output is filtered to retain the UNL relations (semantic relations) only which is shown in Table VII.

TABLE VII
UNL EXPRESSION FOR RTE-3 TEST ANNOTATED SET HYPOTHESIS

```
aoj(serve, kennon)
obj(serve, justice)
```

2 http://unl.ru/deco.html

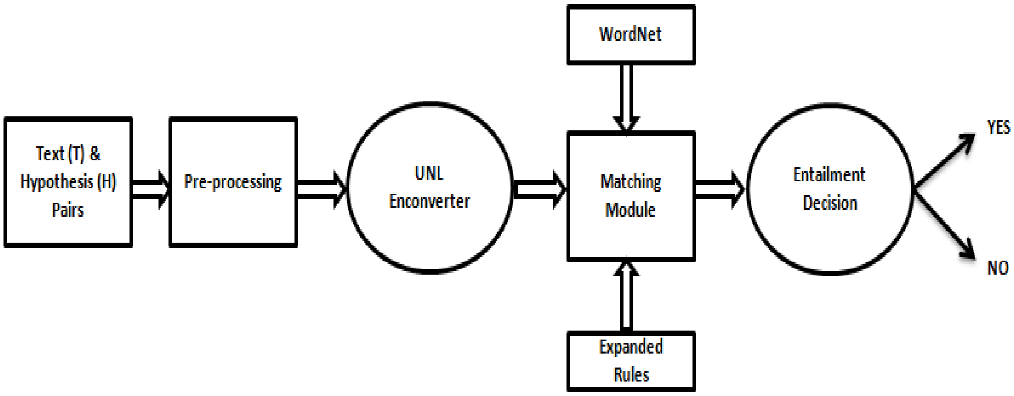

Fig. 1. Semantic Textual Entailment System.

C. Matching Module

After UNL relations are identified for both the text and the hypothesis in each pair, the hypothesis UNL relations are compared with the text UNL relations. The different features that are compared are explained below. In all comparisons, a matching score of 1 is considered when the complete UNL relation along with all of its arguments matches in both the text and the hypothesis. In case of a partial match for a UNL relation, a matching score of 0.5 is assumed. We used the partial match in Rule 3 only.

TABLE VIII
UNL EXPRESSION

Previous Relation	Expand Relation	Example
mod(x,y)	aoj(y,x)	*Red Leaf ⇒ Leaf is Red*
pos(x,y)	mod(y,x)	*Newton's Law ⇒ Newton Law*
aoj(x,y), aoj(y,z)	aoj(x,z)	*He is a boy. A boy is a man. ⇒ He is a man.*
pos(x,y), agt(z,x)	agt(z,y)	*Chief Minister of West Bengal said the thing. ⇒ West Bengal said the thing.*
man(x,y),bas(x,y)	aoj(x,z)	*A rose is more beautiful than tulip. ⇒ Rose is beautiful.*
ins(x,y)	ins(x,z), if z is a hypernym of y.	*He sang with a guitar. ⇒ He sang with an instrument.*
pos(x,y)	iof(x,y)	*Tokyo is a city in Japan. ⇒ Tokyo is a city of Japan.*
and(x,y),and(y,z)	and(y,z)	*You and me., Me and Ramesh. ⇒ Ramesh and you.*

Rule 1: Match Relation = (Number of hypothesis UNL relations that match with text / Number of hypothesis UNL relations)

If Match Relation is above 60%, then this pair is marked as "YES", otherwise as "NO".

Rule 2: If the above Match Relation entailment value is "NO" then we apply the expanded rule given below in both the hypothesis and the text file.

Match Relation (Expand rule) = (Number of hypothesis UNL relations that match with text (obtained from Rule 1) + Number of hypothesis UNL relations that match with text by Expand rule / Number of hypothesis UNL relations).

Expand rules are applicable to those UNL relations that do not match in Rule 1. If Match Relation (Expand rule) is above 60%, then this pair is marked as "YES", otherwise as "NO".

Rule 3: If Match Relation (Expand rule) entailment value is "NO" then we apply the Rule 3 as given below in both the hypothesis and the text file.

Match Relation (Partial Expand rule) = (Number of hypothesis UNL relations that match with text (obtained from Rule 1) + Number of hypothesis UNL relations that match with text by Expand rule (obtained from Rule 2) + Number of hypothesis UNL relation match with text by WordNet synonym / Number of hypothesis UNL relations).

If Match Relation (Partial Expand rule) is above 60% then this pair marked as "YES", otherwise as "NO".

IV. EXPERIMENTAL RESULTS

In RTE-4, no development set was provided, as the pairs proposed were very similar to the ones contained in RTE-3 development and test sets, which could therefore be used to train the systems. Four applications—namely IE, IR, QA and SUM—were considered as settings or contexts for the pair generation. The length of the H's was the same as in the past data sets (RTE-3); however, the T's were generally longer. The RTE-3 test annotated set was used to train our entailment system to identify the threshold values for the various measures towards entailment decision. The two-way RTE-3 test annotated set consisted of 800 text–hypothesis pairs. The RTE-4 test set consisted of 1000 text–hypothesis pairs.

Two baseline systems have been developed in the present task. The Baseline-1 system assigns YES tag to all the text-hypothesis pairs and the Baseline-2 system assigns NO tag to all the text hypothesis pairs.

TABLE IX
BASELINE SYSTEMS FOR RTE-3 DEVELOPMENT SET AND RTE-4 TEST SET:
STANDS FOR THE NUMBER OF DECISIONS, P FOR PRECISION

	Decision	Gold standard	Baseline-1		Baseline-2	
			#	P, %	#	P, %
RTE-3 Development Set	YES	410	800	51.25	0	0
	NO	390	0	0	800	48.75
RTE-4 Test Set	YES	500	1000	50.00	0	0
	NO	500	0	0	1000	50.00

The results obtained on Baseline-1 and Baseline-2 systems on the RTE-3 development data set and the RTE-4 test data set are shown in Table IX.

In our textual entailment system, the method was run separately on the RTE-3 test annotated set and two-way entailment (“YES” or “NO”) decisions were obtained for each text-hypothesis pair. Experiments were carried out to measure the performance of the final RTE system. It is observed that the precision and recall measures of the final RTE system are best when final entailment decision is based on entailment value (YES/NO) results with threshold value 0.60. The results on the RTE-3 test annotated data set are shown in Table X.

TABLE X
UNL RTE-3 DEVELOPMENT SET STATISTICS FOR OUR SYSTEM WITH DIFFERENT THRESHOLD VALUES

		Threshold		
		0.50	0.60	0.70
“YES”	System	572	481	461
	System ∩ Gold	313	278	257
	Gold	410	410	410
	Precision, %	54.72	57.79	55.74
	Recall, %	76.34	67.80	62.68
“NO”	System	228	319	339
	System ∩ Gold	131	204	186
	Gold	390	390	390
	Precision, %	57.45	63.94	54.86
	Recall, %	33.58	52.30	47.69

Experiments were carried out to measure the performance of the final RTE system. The results on the RTE-3 test annotated set for "YES" and "NO" entailment decisions are shown in Table XI.

TABLE XI
RTE-3 TEST ANNOTATED DATA SET STATISTICS FOR OUR SYSTEM, WITH THRESHOLD VALUE 0.60

Entailment Decision	Gold standard	System, correct	System, total	Precision	Recall
YES	410	278	481	57.79%	67.80%
NO	390	204	319	63.94%	52.30%
Total	800	482	800	60.25%	60.25%

The results on RTE-4 test set are shown in Table XII.

TABLE XII
RTE-4 TEST SET STATISTICS FOR OUR SYSTEM, WITH THRESHOLD VALUE 0.60

Entailment Decision	Gold standard	System, correct	System, total	Precision	Recall
YES	500	327	585	55.89%	65.40%
NO	500	276	415	66.50%	55.20%
OVERALL	1000	603	1000	60.30%	60.30%

V. CONCLUSION

Our results show that a Semantic-based approach appropriately tackles the textual entailment problem. Experiments have been initiated for a semantic and syntactic based RTE task.

The next step is to carry out detailed error analysis of the present system and identify ways to overcome the errors. In the present task, the final RTE system has been optimized for the entailment YES/NO decision using the development set. The role of the application setting for the RTE task has also not yet been looked into. This needs to be experimented in the future. The two-way task has to be upgraded to the three-way task.

Finally, given that graph-matching is a computationally expensive task [10], we plan to improve the computational efficiency of our algorithm.

REFERENCES

[1] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, “The Second PASCAL Recognising Textual Entailment Challenge,” in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy. 2006.

[2] L. Bentivogli, I. Dagan, H.T. Dang, D. Giampiccolo, and B. Magnini, “The Fifth PASCAL Recognizing Textual Entailment Challenge,” in *TAC 2009 Workshop*, National Institute of Standards and Technology Gaithersburg, Maryland USA. 2009.

[3] J. Cardeñosa, A. Gelbukh, E. Tovar (eds.), *Universal Networking Language: Advances in Theory and Applications*, IPN. 2005; www.gelbukh.com/UNL-book.

[4] I. Dagan, O. Glickman, and B. Magnini, “The PASCAL Recognising Textual Entailment Challenge,” in *Proceedings of the PASCAL RTE Challenge*. 2005.

[5] D. Giampiccolo, H. T. Dang, B. Magnini, I. Dagan, and E. Cabrio, “The Fourth PASCAL Recognizing Textual Entailment Challenge,” in *TAC 2008 Proceedings*. 2008.

[6] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, “The Third PASCAL Recognizing Textual Entailment Challenge,” in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. 2007.

[7] UNDL Foundation, *Universal networking language (UNL) specifications*, edition 2006, August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/.

[8] E.M. Voorhees and D. Harman, “Overview of the Seventh Text Retrieval Conference,” in *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication. 1999.

[9] D. Yuret, A. Han, and Z. Turgut, “SemEval-2010 Task 12: Parser Evaluation using Textual Entailments,” in *Proceedings of the SemEval-2010 Evaluation Exercises on Semantic Evaluation*. 2010.

[10] G. Zhao, M. Petridis, G. Sidorov, and J. Ma, “A critical examination of node similarity graph matching algorithm,” *Research in computing science*, vol. 40, pp. 73–82, 2008.