THK's Natural Logic-based Compositional Textual Entailment Model at NTCIR-10 RITE-2

Yotaro Watanabe
Tohoku University, Japan
yotaro-w@ecei.tohoku.ac.jp

Junta Mizuno
Tohoku University, Japan
junta-m@ecei.tohoku.ac.jp

Kentaro Inui
Tohoku University, Japan
inui@ecei.tohoku.ac.jp

ABSTRACT

This paper describes the THK system that participated in the BC subtask, MC subtask, ExamBC subtask and UnitTest in NTCIR-10 RITE-2. Our system learns plausible transformations of pairs of *Text* t_1 and *Hypothesis* t_2 only from semantic labels of the pairs using a discriminative probabilistic model combined with the framework of *Natural Logic*. The model is trained so as to prefer alignments and their semantic relations which infer the correct sentence-level semantic relations. In the formal run, we achieved the highest performance of detecting contradictions in the MC subtask (28.57 of F1).

Team Name

Tohoku University / THK

Subtasks

BC subtask, MC subtask, ExamBC subtask and UnitTest (Japanese)

Keywords

textual entailment, natural logic, log-linear model, hidden variables

1. INTRODUCTION

In this paper, we describe the THK system, a Japanese textual entailment recognition system, that participated in the NTCIR-10 RITE-2 (Recognizing Inference in TExt) [9]. Our system is based on the latent discriminative model combined with the framework of Natural Logic proposed by Watanabe et al. [10]. The model aligns units in two sentences t_1 and t_2, and labels semantic relations defined in Natural Logic [4] to the units. Then a sentence-level semantic relation is inferred from the set of alignments, their semantic relations and their contexts. The parameters of the model is trained so as to prefer alignments that infer correct sentence-level semantic relations. We report the results of our system that participated in the four subtasks: BC subtask, MC subtask, ExamBC subtask and UnitTest. In the MC subtask, our system achieved the highest performance in terms of contradiction relation detection.

This paper is organized as follows. At first we describe the theory of Natural Logic (Section 2) and an overview of our system (Section 3). Then we report the setting for the formal run (Section 4) and the results (Section 5). Finally we conclude in Section 6.

2. NATURAL LOGIC

The concept of *Natural Logic*, a logic over natural language, is originally proposed by [1], and then [7, 8] and [6] explored monotonicity calculus[1] to explain entailment relations using Natural Logic. While they considered only containment relations, [4] introduced an *exclusion* relation to deal with entailment relations which involve different objects or concepts (e.g. Stimpy is a cat $\models$ Stimpy is not a poodle). In this section, we describe the theory of Natural Logic proposed by [4, 2].

The basic idea of MacCartney et al's theory is that the semantic relation between sentences can be derived from the semantic relations of *edits* (substitution, deletion and insertion) from T to H. The fundamental assumption of the theory is *compositionality*: (some of) the entailments of a compound expression are a function of the entailments of its parts. They defined the seven types of semantic relations for edits: equivalence ($a \equiv b$ if $a = b$), forward-entailment ($a \sqsubset b$ if $a \subset b$), backward-entailment ($a \sqsupset b$ if $b \supset a$), negation ($a \ ^\wedge\ b$ if $a \cap b = \phi \wedge a \cup b = U$)[2], alternation ($a \mid b$ if $a \cap b = \phi \wedge a \cup b \neq U$), cover ($a {}^{\cup} b$ if $a \cup b \neq \phi \wedge a \cup b = U$), and independence ($a \# b$ otherwise).

Semantic relations provided by edits are *projected* onto other relations depending on their contexts using *projection rules*. For example, in a scope of negation, forward-entailment is projected onto backward-entailment (e.g *soccer* $\sqsubset$ *sports*, *I didn't play soccer.* $\sqsupset$ *I didn't play sports.*). Other linguistic expressions such as logical connectives and quantifiers also projects semantic relations. A semantic relation between sentences is derived by combining the projected semantic relations of edits using *composition rules*. The rules are defined as tuples of semantic relations. Let the seven types of relations be $\mathcal{R}$, $r_i \in \mathcal{R}$, $r_j \in \mathcal{R}$, then a compositional rule is represented by $r_i \bowtie r_j \Rightarrow \boldsymbol{r} \subseteq \mathcal{R}$. Some compositional rule derive a single relations (e.g. $\equiv \bowtie \sqsubset \Rightarrow \sqsubset$), and others derive more than one semantic relations (e.g. $\mid \bowtie \mid \Rightarrow \bigcup \{\equiv, \sqsubset, \sqsupset, \mid, \#\}$). As semantic relation composition proceeded, semantic relations tend to move toward $\#$ [3].

[1]In an *upward-monotone* context, replacing a linguistic expression with a more general expression preserves truth. On the other hand, in a *downward-monotone* context, replacing a linguistic expression with a more specific expression preserves truth.
[2]U denotes a universe.
[3]Due to spacial limitations, we can not give all of the composition rules. For more details, see [2].

3. A LATENT DISCRIMINATIVE MODEL WITH NATURAL LOGIC

Our approach used for the subtasks in RITE-2 is based on the model proposed by Watanabe et al. [10]. The model assigns local semantic relations to edits which represent a transformation from T to H. A valid set of edits represents an alignment between T and H. Each edit is categorized as one of three types: *substitution*, *deletion* or *insertion*, and is given one of the seven semantic relations defined in Natural Logic described in Section 2. A semantic relation between T and H is derived from a set of semantic relations of alignment edits by using the projection rules and the composition rules.

The model learns appropriate alignments which are consistent with compositional rules of Natural Logic from only sentence-level semantic relations, where appropriate alignments, their semantic relations and their projections are represented using hidden variables. We use a log-linear discriminative model with hidden variables to provide conditional joint probabilities of alignments, their associated semantic relations, and their projections and a sentence-level semantic relation.

3.1 Model

The model provides a conditional joint distribution of alignment edits, their semantic relations, their projected relations and the final semantic relation between T and H as follows.

$$p(\boldsymbol{e}, \boldsymbol{r_e}, \boldsymbol{r_e^P}, \boldsymbol{r}^C | \boldsymbol{x}; \boldsymbol{\lambda}) \propto \exp\left(\sum_k \Psi_k(\boldsymbol{e}, \boldsymbol{r_e}, \boldsymbol{r_e^P}, \boldsymbol{r}^C, \boldsymbol{x}; \boldsymbol{\lambda})\right) \quad (1)$$

where $\boldsymbol{e} = \{e_i\}$ denotes the variables representing edits, and each edit $e_i = \langle \boldsymbol{t}_i, \boldsymbol{h}_i \rangle$ consists of $\boldsymbol{t}_i$, a subset of indices of units (e.g. words) in T, and $\boldsymbol{h}_i$, a subset of indices of units in H. An edit corresponds to substitution if $\boldsymbol{t}_i \neq \phi$ and $\boldsymbol{h}_i \neq \phi$, deletion if $\boldsymbol{t}_i \neq \phi$ and $\boldsymbol{h}_i = \phi$, and insertion if $\boldsymbol{t}_i = \phi$ and $\boldsymbol{h}_i \neq \phi$. $\boldsymbol{r_e}$ represents the set of semantic relations for $\boldsymbol{e}$, where $r_{e_i} \in \boldsymbol{r_e}$ corresponds to the semantic relation of e_i. Since r_{e_i} is derived without considering its context, r_{e_i} can be seen as the semantic relation between $\boldsymbol{t}_i$ and $\boldsymbol{h}_i$. The variables $\boldsymbol{r_e^P}$ represents a set of projected semantic relations derived from $r_{\boldsymbol{e}}$, taking into account their contexts. If an edit is under the scope of negation, a quantifier or a conditional, then r_{e_i} is mapped to an appropriate semantic relation $r_{e_i}^P$ based on that context. Therefore $r_{e_i}^P$ can be seen as the sentence-level semantic relation between T and the sentence which can be obtained by applying the edit e_i to T. The variables $\boldsymbol{r}^C$ denotes a set of semantic relations derived by combining $\boldsymbol{r_e^P}$, where each $r^C \in \boldsymbol{r}^C$ corresponds to the result of composition of two semantic relations. Hereafter, we use r_T^C as the sentence-level semantic relation. Note that $r_T^C \in \boldsymbol{r}^C$. Each variable r in $\boldsymbol{r_e}$, $\boldsymbol{r_e^P}$ and $\boldsymbol{r}^C$ can have seven types of semantic relations described previously. Ψ_k in equation (1) is a *factor* which scores the plausibility of alignment edits, their semantic relations, etc.

The model uses the following four types of factors to score the plausible alignment edits, their semantic relations and a sentence-level semantic relation.

Alignment Factor $\Psi_A(e, \boldsymbol{x})$ is used to deal with (unlabeled) phrase alignment for entailment relation recognition and is defined as $\Psi_A(e, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}_A(e, \boldsymbol{x})$. In order to provide good alignments, it is necessary to capture the lexical similarity between words. The features used in this factor are mainly (i) surface-based similarity between alignment units, (ii) semantic relatedness of alignment units, which can be extracted from diverse lexical knowledge databases, and (iii) the contextual information for an edit.

Alignment Semantic Relation Factor $\Psi_S(e, r_e, \boldsymbol{x})$ is introduced to provide plausibility of a semantic relation $r_e \in \boldsymbol{r_e}$ for an alignment edit $e \in \boldsymbol{e}$ and is defined as $\Psi_S = \boldsymbol{\lambda} \cdot \boldsymbol{f}_S(e, r_e, \boldsymbol{x})$. Each variable r_e has a distribution over the seven types of semantic relations defined in Natural Logic. In order to classify semantic relations, not only surface-based similarities, but also lexical semantic relations play an important role. In the NatLog system developed by [2], an implementation of an RTE system of Natural Logic, lexical resource-derived features (e.g. WordNet, NomBank, etc.), string similarity features, and lexical category features are used. For this factor, we exploit diverse lexical resources to provide informative features for classifying semantic relations of edits.

Projection Factor $\Psi_P(r_e, r_e^P, \boldsymbol{x})$ provides an appropriate projection from r_e to r_e^P by considering the context of e, and is defined by $\Psi_P(r_e, r_e^P, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}_P(r_e, r_e^P, \boldsymbol{x})$. This factor captures the effects of monotonicity (e.g. upward, downward). Given r_e and its contexts, the semantic relation of the projected variable r_e^P is uniquely determined using the monotonicity rules of [2].

Composition Factor $\Psi_C(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x})$ scores tuples of semantic relations, and is defined by $\Psi_{C_i}(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x})$. In this factor, we use the composition rules used in [2] with some modification. We set the derived semantic relations to independence (#) for the rules which derive more than one semantic relations. Therefore, as with Ψ_P, given two semantic relations r_{i-1}^C and r_e^P, the joined relation of the variable r_i^C is uniquely determined.

An overview of the model is shown in Figure 1. In this figure, we show the factor graph constructed by the model for a pair of sentences in Japanese. Our model is divided by three layers: the *alignment layer*, the *projection layer* and the *composition layer*. First, in the alignment layer, the model scores possible alignments using Ψ_A and Ψ_S. For alignment units, we use *bunsetsu* which is a reasonable unit for Japanese linguistic analysis. A *bunsetsu* is a chunk-like unit that consists of one or more content words and zero or more functional words. A set of possible alignments are obtained using an extended MANLI algorithm [3].

Next, for each alignment obtained by the alignment algorithm, we construct a factor graph as shown in Figure 1. The factor graph has variables for alignments, projected relations, joined relations, and the factors defined previously. In the projection layer, semantic relations of alignments are projected by Ψ_P, and finally a sentence-level semantic relation is obtained in the composition layer using the projected relations and composition rules encoded in Ψ_C.

In inference, since variables related to Ψ_P and Ψ_C are uniquely determined if $\boldsymbol{r_e}$ is given, the model derives the best

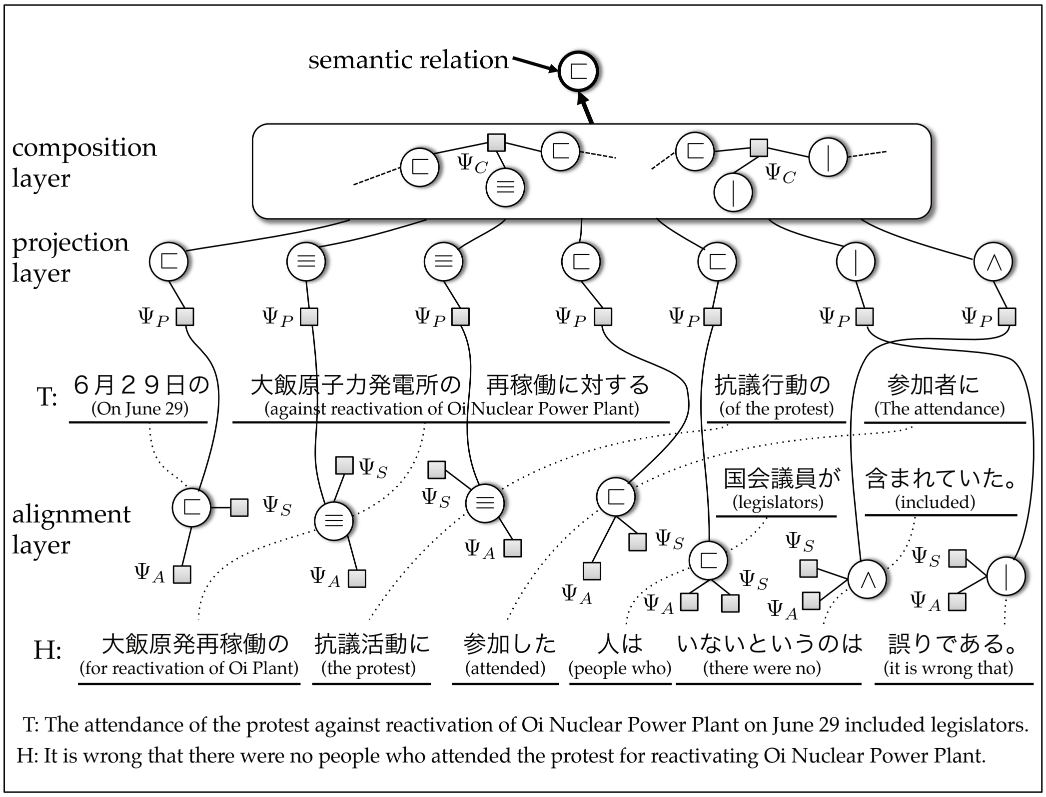

Figure 1: An overview of the model.

alignments, their semantic relations, and a sentence-level semantic relation simultaneously. In training, the parameters of the model are updated so as to derive alignments and their semantic relations which derive the correct sentence-level semantic relation based on the composition rules of Natural Logic.

3.2 Learning and Inference of the Model

The parameters $\boldsymbol{\lambda}$ of the model were trained from sentence-level semantic relations via marginal-likelihood maximization $\mathcal{L}_{\boldsymbol{\lambda}} = \sum_n \log p(r_T^C = l^n | \boldsymbol{x}^n; \boldsymbol{\lambda})$.

The partial differential of the objective function is

$$\frac{\partial L}{\partial \lambda_k} = \sum_n \left(\frac{\partial}{\partial \lambda_k} \log \sum_{\langle \boldsymbol{e}, \boldsymbol{r_e} \rangle \in \mathcal{E}} \sum_{\boldsymbol{r}: r_T^C = l} \exp\left(\sum_k \boldsymbol{\Psi}_k(\boldsymbol{e}, \boldsymbol{r_e}, \boldsymbol{r_e^P}, \boldsymbol{r}^C, \boldsymbol{x}) \right) - \frac{\partial}{\partial \lambda_k} \log \widetilde{Z}(\boldsymbol{x}) \right) \quad (2)$$

where edits $\boldsymbol{e}$, their semantic relations $\boldsymbol{r_e}$, projected semantic relations $\boldsymbol{r_e^P}$ and joined relations excluding the sentence-level semantic relation are all hidden variables.

In optimization, only the parameters in Ψ_A and Ψ_S are updated, and the parameters in Ψ_P and Ψ_C are left to initial values. In order to update the parameters, we need to calculate marginal probabilities of the alignments. We use only N-bests provided by the extended MANLI algorithm proposed by [10] to calculate approximate marginal probabilities which are derived from an approximate partition function $\widetilde{Z}(\boldsymbol{x})$ instead of the exact partition function $Z(\boldsymbol{x})$,

The inference phase also uses the extended MANLI algorithm to estimate the most plausible assignments of alignments, their semantic relations, their projections and their compositions.

We need to convert the NL relations to corresponding relation in binary and multiclass subtasks in RITE-2. In training and inference of the model, for the BC subtask, ExamBC subtask and UnitTest, $\equiv$ and $\sqsubset$ is converted to Y and N otherwise ($\sqsupset$, $^\wedge$, $|$, $^\cup$ and $\#$). For the MC subtask, $\equiv$ is mapped to B, $\sqsubset$ is mapped to F, $^\wedge$ and $|$ are mapped to C, and $^\cup$ and $\#$ are mapped to I.

4. SETTINGS

For the factors Ψ_P and Ψ_C, we initialized the weights to 0.0 if the semantic relation tuple is covered by our projection rules and composition rules, and $-\infty$ otherwise. For the factors Ψ_A and Ψ_S, we set initial weights to some features. In training of the model, we update the parameters in Ψ_A and Ψ_S, and the parameters in Ψ_P and Ψ_C are left to the initial values. Parameter updating was performed using stochastic gradient descent (SGD), and the number of iterations was set to 2. Also, we applied L_2 regularization. As for the alignment algorithm, the number of iterations was set to 40, and the number of N-bests was set to 10. For each edit type, we restricted the maximum size of units: only allows one-to-one for substitution, allows at most three units for insertion and deletion edits. Also, we constrained the types of semantic relations for each edit type. Substitution edits can have one of the five types of semantic relations: $\equiv$, $\sqsubset$, $\sqsupset$, $^\wedge$ and $|$ with an exception. If the lemma sequences of the

BC	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
THK-02 (unsubmitted)	58.34	58.69	62.16	50.49	80.86	54.51	75.50	42.66
THK-01	52.40	53.28	45.92	44.65	47.27	58.87	60.18	57.63
Baseline-01	62.53	63.93	55.28	57.63	53.13	69.78	67.91	71.75
Top	80.49	81.64	75.76	84.95	68.36	85.22	79.95	91.24
ExamBC	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
THK-02 (unsubmitted)	46.59	46.65	48.38	38.62	64.74	44.80	61.39	35.27
THK-01	43.77	62.28	11.52	61.11	6.36	76.03	62.33	97.45
Baseline	54.77	56.47	45.98	44.15	47.98	63.55	65.38	61.82
Top	67.15	70.31	56.96	64.71	50.87	77.34	72.76	82.55
UnitTest	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
THK-02 (unsubmitted)	56.59	73.86	83.93	91.16	77.83	29.21	21.67	44.83
THK-01	53.26	71.37	82.35	89.94	75.94	24.18	17.74	37.93
Baseline	51.70	86.31	92.58	88.41	97.17	10.81	25.00	6.90
Top	77.77	90.87	94.84	94.39	95.28	60.71	62.96	58.62

Table 1: Results on the BC subtask, ExamBC subtask and UnitTest formal run data (JA).

two *bunsetsus* are the same, the edit can have only $\equiv$. Deletion edits and insertion edits can have $\sqsubset$ and $\sqsupset$ respectively with exceptions. They can have $|$ if the head of *bunsetsu* matches an entry in the list of *counter-factive expressions* [4], and they can have $\equiv$ if the head of *bunsetsu* matches an entry in the list of *less-informative expressions* [5].

The features used in the model are the same as those used in [10].

5. RESULTS

We submitted only one run for each subtask (denoted as THK-01). In the THK-01 setting, for each subtask, the model was trained with the corresponding development data of each subtask. In this paper, we additionally report the results (denoted as THK-02) where the model was trained with the MC-dev data and tested on the formal run data of the binary-class subtasks (BC subtask, ExamBC subtask and UnitTest). Since more fine-grained semantic relations provide more effective information for estimating plausible alignments, we expect the model to achieve performance improvements.

Table 1 shows the results on the dataset of the binary-class subtasks (BC subtask, ExamBC subtask and UnitTest). We achieved slightly better performances compared to a machine learning-based baseline in UnitTest, however defeated in the BC subtask and the ExamBC subtask. The reason of the insufficient performance is that we did not include shallow word overlapping features in the model. Simple word overlap features are known to have strong correlations with entailment relations (e.g. [5]).

As described in Section 3, we regarded both $\equiv$ and $\sqsubset$ as correct if a pair of sentence has an entailment relation (i.e. "Y"). However, in this case, only the "Y" label does not provide sufficient information to discriminate between *equivalence* and *entailment*. Since the examples of the MC dataset have more fine-grained semantic labels, we expect our model to achieve performance improvements. The results of THK-02 in the table suggest that more fine-grained semantic relations are effective for estimating better parameters in alignment.

Table 2 shows the result on the MC formal run data. Although the performance of our model is insufficient compared to the top system, our model achieved the highest performance on contradiction detection. We found out that the model assigned $^\wedge$ or $|$ relation to alignments which have a unit with negation expression such as *nai* with higher recall. Since our strategy of the compositional model combines local semantic relations of alignments, the alignments that have *opposite* relation such as $^\wedge$ and $|$ guided the model to infer contradiction relations successfully.

Table 4 shows the confusion matrix of the results on the MC subtask formal run data. Since the system outputs not only forward-entailment relations, but also "reverse" entailment relations, "R" is included in the table. Most of the pairs of bi-directional entailment fell into the forward-entailment relation because the model failed to assign the $\equiv$ relation to delete or insert meaning-less expressions such as ⟨ 主に *omoni* "mainly"⟩, ⟨俗に *zokuni* "commonly"⟩, ⟨一種 *isshu* "a kind of"⟩, etc. Since our model assigns forward-entailment to deletion edits and reverse-entailment to insertion edits by default, the results left from bi-directional entailment.

The model tends to frequently output forward-entailment relations due to the training strategy of the model. In training, all of sets of alignments which infer forward-entailment are regarded as true if the label of an examples is forward-entailment, and there are many possible alignments in such cases. As a result, the model trained by such strategy can easily output forward-entailment relations of alignments even for very dissimilar units.

gold \system	B	F	R	C	I	Total
B	9	32	8	1	20	70
F	0	171	1	2	31	205
C	1	32	3	12	13	61
I	2	123	2	8	77	212
Total	12	358	14	23	141	548

Table 4: Confusion Matrix on the MC subtask data (JA).

[4]A hand-crafted list which contains 13 entries.

[5]As with the list of *counter-factive expressions*, the list was hand-crafted, and contains 30 entries.

MC	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
THK-01	30.98	49.09	21.95	75.00	12.86	60.75	47.77	83.41	28.57	52.17	19.67	43.63	54.61	36.32
Baseline	26.61	45.44	0.00	0.00	0.00	56.18	43.01	80.98	5.41	15.38	3.28	44.88	54.36	38.21
Top	59.96	69.53	67.18	72.13	62.86	76.47	76.85	76.10	21.15	25.58	18.03	75.06	70.54	80.19

Table 2: Results on the MC subtask formal run data (JA).

Category	Y:Prec		Y:Rec		Y:F1	N:Prec		N:Rec		N:F1
case_alternation	100.00	(7/ 7)	100.00	(7/ 7)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
inference	0.00	(0/ 0)	0.00	(0/ 2)	0.00	0.00	(0/ 2)	0.00	(0/ 0)	0.00
spatial	100.00	(1/ 1)	100.00	(1/ 1)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
implicit_relation	100.00	(3/ 3)	16.67	(3/ 18)	28.57	0.00	(0/ 15)	0.00	(0/ 0)	0.00
list	100.00	(3/ 3)	100.00	(3/ 3)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
disagree:lex	0.00	(0/ 1)	0.00	(0/ 0)	0.00	100.00	(1/ 1)	50.00	(1/ 2)	66.67
synonymy:phrase	100.00	(24/ 24)	68.57	(24/ 35)	81.36	0.00	(0/ 11)	0.00	(0/ 0)	0.00
meronymy:lex	100.00	(1/ 1)	100.00	(1/ 1)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
apposition	0.00	(0/ 0)	0.00	(0/ 1)	0.00	0.00	(0/ 1)	0.00	(0/ 0)	0.00
modifier	100.00	(40/ 40)	95.24	(40/ 42)	97.56	0.00	(0/ 2)	0.00	(0/ 0)	0.00
transparent_head	0.00	(0/ 0)	0.00	(0/ 1)	0.00	0.00	(0/ 1)	0.00	(0/ 0)	0.00
synonymy:lex	100.00	(10/ 10)	100.00	(10/ 10)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
nominalization	0.00	(0/ 0)	0.00	(0/ 1)	0.00	0.00	(0/ 1)	0.00	(0/ 0)	0.00
coreference	0.00	(0/ 0)	0.00	(0/ 4)	0.00	0.00	(0/ 4)	0.00	(0/ 0)	0.00
disagree:phrase	0.00	(0/ 14)	0.00	(0/ 0)	0.00	100.00	(11/ 11)	44.00	(11/ 25)	61.11
temporal	100.00	(1/ 1)	100.00	(1/ 1)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
disagree:modality	0.00	(0/ 1)	0.00	(0/ 0)	0.00	0.00	(0/ 0)	0.00	(0/ 1)	0.00
entailment:phrase	100.00	(38/ 38)	84.44	(38/ 45)	91.57	0.00	(0/ 7)	0.00	(0/ 0)	0.00
disagree:temporal	0.00	(0/ 0)	0.00	(0/ 0)	0.00	100.00	(1/ 1)	100.00	(1/ 1)	100.00
hypernymy:lex	100.00	(3/ 3)	100.00	(3/ 3)	100.00	0.00	(0/ 0)	0.00	(0/ 0)	0.00
scrambling	100.00	(14/ 14)	93.33	(14/ 15)	96.55	0.00	(0/ 1)	0.00	(0/ 0)	0.00
clause	100.00	(13/ 13)	92.86	(13/ 14)	96.30	0.00	(0/ 1)	0.00	(0/ 0)	0.00
relative_clause	100.00	(7/ 7)	87.50	(7/ 8)	93.33	0.00	(0/ 1)	0.00	(0/ 0)	0.00

Table 3: Detailed results on UnitTest data for each category.

Table shows the detailed results on UnitTest data. The proposed model could correctly answer subsets of the examples categorized into "disagree_lex", "disagree_temporal" and "disagree_phrase". However, overall, there are many false positives of "N" examples especially in the categories of "implicit_relation", "synonymy_phrase" and "entailment_phrase". In the "implicit_relation" examples, t_2 has some specific information compared to t_1, however, the information can be complemented with some information in t_1. Since the model has no mechanism of dealing with such complementation, many examples are misclassified. Many examples of synonymy_phrase and entailment_phrase are misclassified because they require phrase alignment, however our alignment algorithm does not support it.

6. CONCLUSION

This paper described the THK system that participated in the BC subtask, MC subtask, ExamBC subtask and UnitTest in NTCIR-10 RITE-2. Our model has ability to predict local correspondences (alignments) between sentences, the semantic relations, and the sentence-level semantic relation simultaneously. The model can be trained from only sentence-level semantic relations by using marginal-likelihood maximization. The results of our model were modest all in all, however the system achieved the best performance in detecting contradiction relations.

7. REFERENCES

[1] G. Lakoff. George Lakoff - Linguistics and Natural Logic. *Synthese*, 22:151–271, 1970.

[2] B. MacCartney. *Natural Language Inference*. PhD thesis, 2009.

[3] B. MacCartney, M. Galley, and C. Manning. A Phrase-Based Alignment Model for Natural Language Inference. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, 2008.

[4] B. MacCartney and C. D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, 2008.

[5] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 291–301, 2011.

[6] V. S. Valencia. *Studies on Natural Logic and Categorial Grammar*. PhD thesis, 1991.

[7] J. van Benthem. The semantics of variety in categorial grammars. pages 33–55, 1988.

[8] J. van Benthem. Language in action: Categories, lambdas and dynamic logic. 1991.

[9] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, N. Kando, H. Shima, and K. Takeda.

Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 2013.

[10] Y. Watanabe, J. Mizuno, E. Nichols, N. Okazaki, and K. Inui. A latent discriminative model for compositional entailment relation recognition using natural logic. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, 2012.