Labeled Alignment for Recognizing Textual Entailment [*]

Xiao-Lin Wang   Hai Zhao   Bao-Liang Lu
Center for Brain-Like Computing and Machine Intelligence
MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai 200240, China
arthur.xl.wang@gmail.com   {zhaohai,blu}@cs.sjtu.edu.cn

Abstract

Recognizing Textual Entailment (RTE) is to predict whether one text fragment can semantically infer another, which is required across multiple applications of natural language processing. The conventional alignment scheme, which is developed for machine translation, only marks the paraphrases and hyponyms to justify the entailment pairs, while provides less support for the non-entailment ones. This paper proposes a novel alignment scheme, named labeled alignment, to address this problem, which introduces negative links to explicitly mark the contradictory expressions to justify the non-entailment pairs. Thus the alignment-based RTE method employing the proposed scheme, compared with those employing the conventional one, can gain accuracy improvement through actively detecting the signals of non-entailment. The experimental results on the data sets of two shared RTE tasks indicate the implemented system significantly outperforms both the baseline system and all the other submitted systems.

1 Introduction

Textual Entailment (TE) is a directional relation between two text fragments. One natural-language premise, noted as P, entails one natural-language hypothesis, noted as H, if typically a human reading P would infer that H is most likely true (Dagan et al., 2006).

Recognizing Textual Entailment (RTE) is proposed as a generic task that captures the semantic inference need across a wide range of natural language processing applications. For example, a question answering system should identify the texts that entail a hypothesized answer, e.g., given the question "*What does Peugeot manufacture?*", the text "*Chrétien visited Peugeot's newly renovated car factory*" entails the hypothesized answer form "*Peugeot manufactures cars*" (Dagan et al., 2006). Similarly, in Machine Translation (MT) evaluation, a correct translation should be semantically equivalent to the gold translation, that is, both translations should entail each other (Padó et al., 2009).

RTE has attracted extensive attention ever since it was proposed. A wide range of methods have been proposed, and quite a few successful approaches treat RTE as an alignment problem. Alignment is originally developed for MT to bridge two languages (Brown et al., 1993). Alignment is to establish links between the semantically equivalent atom expressions in two sentences. (Marsi and Krahmer, 2005) first advocates pipelined system architectures that contain distinct alignment components. This latter becomes a strategy crucial to the top-performing systems of (Hickl et al., 2006). In addition, human-generated alignment annotations for the second PASCAL [1] RTE challenge is released by Microsoft Research to facilitate related research (Brockett, 2007).

The principle of the existing alignment-based RTE methods is that a sufficiently good alignment between P and H means a close lexical and structural correspondence, thus P probably entails H. For example, Fig. (1a) shows that the entailment

---

[1]PASCAL is the European Commission's ICT-funded Network of Excellence for Cognitive Systems, Interaction & Robotics.

relation can be correctly predicted through recognizing "*read into*" → "*interpreted*"[2] and "*what he wanted*" → "*in his own way*".

However, the alignment developed in MT does not solve the non-alignment samples well. It usually links the words in H, which have no counterparts in P, to *NULL* regardless their impacts on the entailment relation. For example, in Fig. (1b), "*ferry sinking*", "*cause*" and "*that*" are all linked to *NULL*[3], while only "*ferry sinking*" is the cause for non-entailment. Thus such an alignment is improper for RTE.

This paper extends the normal alignment scheme to meet the challenge of RTE. The proposed scheme, named labeled alignment, introduce another type of links, named negative links, to mark those critical RTE-related linguistic phenomena that cannot be captured by the normal alignment. For example, Fig. (1c) shows that the previous vital expressions "*ferry sinking*" is linked to "*flood*" through a negative link, noted as "*ferry sinking*" $\nrightarrow$ "*flood*".

The proposed labeled alignment, which explicitly marks the causes of non-entailment, can facilitate the design of RTE method. This paper proposes an RTE method based on the labeled alignments that actively looks for the signal of negative links in order to correctly recall non-entailment samples.

The main contributions of this paper are as follows,

- A labeled alignment scheme is proposed for RTE;

- An RTE data set annotated with the proposed scheme is released;

- High prediction accuracies are achieved on two RTE data sets.

2 Related Work

RTE has attracted extensive attention in the past decade, and a wide range of approaches have been proposed besides the alignment-based methods (Androutsopoulos and Malakasiotis, 2009). The logic-based methods interpret sentences to first-order-logic expressions and then invoke theorem provers (Bos and Markert, 2005). Similarity-based methods employ classifiers to learn from

multiple similarity measures including lexical similarities (Watanabe et al., 2012),edit distance (Rios and Gelbukh, 2012), measurements from MT (Volokh and Neumann, 2011), syntactic tree similarity (Mehdad, 2009) and dependency similarity (Wang and Zhang, 2009). Transform-based methods take entailment as finding a credible transform from the premise to the hypothesis (Kouylekov et al., 2011).

(MacCartney et al., 2008) argues the alignment techniques and tools for MT such as GIZA++ (Och and Ney, 2003) do not readily transfer to RTE. They compare the alignment for RTE with that for MT, and state the following differences:

- The alignment for RTE is monolingual rather than cross-lingual, opening the door to utilizing abundant monolingual resources on semantic relatedness.

- The alignment for RTE is asymmetric, since P is often much longer than H.

- One cannot assume approximate semantic equivalence, since P might be contradictory or independent with H.

- Little training data is available.

They propose a new alignment tool named MANLI for RTE, but still adopts a alignment scheme similar with the one in MT (Brockett, 2007). This paper, however, revises the alignment scheme to support RTE, especially to address the third difference.

(MacCartney et al., 2006) argues that some critical RTE-related linguistic phenomena such as negations and modalities cannot be captured by alignment. They propose a wide range of features to represent them, and employ a classifier to learn from these specialized features as well as the alignment features to predict the entailment relation. The proposed labeled alignment in this paper, however, can natively process these phenomena, e.g., Fig. (2g) solves negations and (2h) solves modalities.

(Sammons et al., 2010) argues that a single label (whether entailment or not) is insufficient to effectively evaluate the performance of RTE system as well as to guide researchers. They raise a group of detailed entailment phenomena such as simple rewriting rules, lexical relations and passive-active transform, as well as a group of detailed

---

[2]The notation means that the expression "*read into*" in P is connected to the expression "*interpreted*" in H.
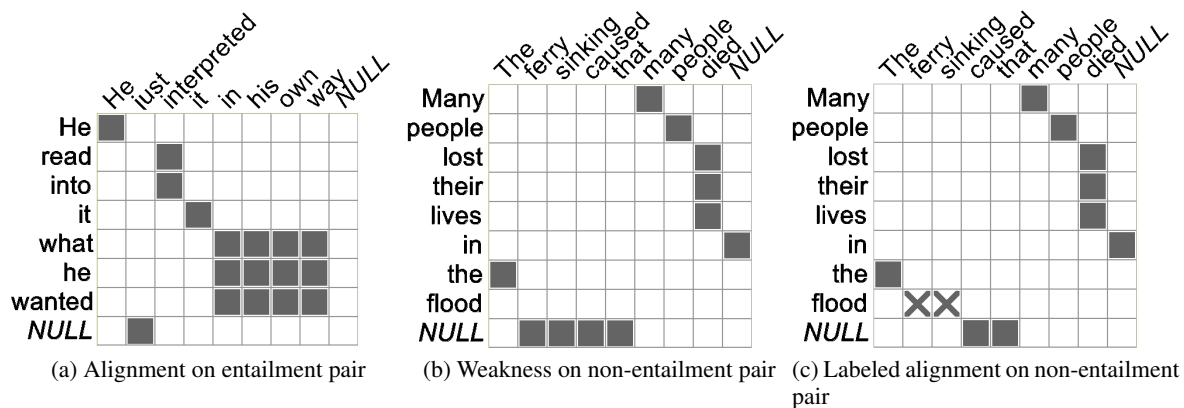
[3]*NULL* means an empty expression.

Figure 1: Illustration of Alignment for RTE. Each subfigure presents an RTE sample. The vertical text is the premise, and the horizontal text is the hypothesis. The solid squares represent positive links, and the crosses represent negative links. (a) is of entailment relation, while (b) and (c) are of non-entailment relations.

non-entailment phenomena such as missing arguments, named entities mismatches and missing modifiers. This paper greatly favors their work, and the proposed labeled alignment scheme can annotate most of the non-entailment phenomena mentioned in their paper, which is beneficial to researchers.

3 Labeled Alignment

Labeled alignment consists of two types of links, named positive link and negative link, respectively. The positive link is inherited from the normal alignment, while the negative link is newly introduced.

3.1 Positive Link

The positive link is inherited from the normal alignment to handle the variability of natural language expressions, that is, the same meaning can be expressed by different texts. The positive link connects the atom expressions e_p in P and e_h in H, if e_p and e_h are paraphrases or e_p infers e_h, noted as $e_p \rightarrow e_h$. As the occurrence of this type of links suggests the entailment relation between P and H, they are named positive links.

This paper partially follows the alignment scheme in (Brockett, 2007; MacCartney et al., 2008) where the links are token-based but many-to-many is allowed, thus multi-word phrases can be explicitly aligned.

The positive links are mainly applied to the following cases:

- identical words;

- synonyms or near synonyms, e.g., *"bought"* → *"purchased"* in Fig. (2a);

- hyponyms, e.g., *"patent"* → *"technology"* in Fig. (2a);

- same named entities, e.g., *"the Microsoft Corporation"* → *"Microsoft"* in Fig. (2a);

- paraphrases or semantically inferable expressions which cannot be further decomposed into smaller links, e.g., *"read into"* → *"interpreted"* and *"what he wanted"* → *"in his own way"* in Fig. (1a);

- trivial words in H versus *NULL*, e.g., *NULL* → *"just"* in Fig. (1a).

3.2 Negative Link

The negative link is introduced to annotate why a RTE sample does not possess an entailment relation. The negative link is noted as $e_p \nrightarrow e_n$ where e_p and e_n are the expressions in P and H, respectively. As the occurrence of this type of links suggests the non-entailment relation, they are named negative links.

The usage of negative links can be divided to three categories – contradictory expressions, unmatched sentence-level modifier and hypothesis novelty.

The contradictory expressions refer to the two expressions from P and H, respectively, which should be compared as motived by the syntactic structures, but actually convey inconsistent semantics. Such phenomena usually lead to the conflic-

tion between P and H. The contradictory expressions include, but are not limited to, the following cases:

- antonyms, e.g., *"catalyst"* $\not\rightarrow$ *"deterrent"* in Fig. (2b);

- mismatches between numbers, dates and times, e.g., *"3 millions"* $\not\rightarrow$ *"10,000"* in Fig. (2c);

- different named entities, e.g., *"Mircrosoft"* $\not\rightarrow$ *"Sony"* in Fig. (2d);

- heads of noun phrases, e.g., *"drill"* $\not\rightarrow$ *NULL* in Fig. (2e);

- vital modifiers of noun phrases, e.g., *"Hispanic"* $\not\rightarrow$ *NULL* in Fig. (2f);

- contradictory content words[4], e.g., *"flood"* $\not\rightarrow$ *"ferry sinking"* in Fig. (1c).

The unmatched sentence-level modifier refers to the modifier in either P or H which impacts the meaning of the whole sentence but has no counterpart in the other sentence. Such phenomena usually flip the entailment relation. The unmatched sentence-level modifier is marked through connecting it to *NULL* through a negative link. The usage includes the following cases:

- negations including simple negation (not), negative quantifiers (no, few), prepositions (without, except), adverbs (never, seldom, nearly), e.g., *"never"* $\not\rightarrow$ *NULL* in Fig. (2g);

- Virtual modalities, e.g., *"could"* $\not\rightarrow$ *NULL* in Fig. (2h);

- phrases that suggest the sentence is not stating a happened event, e.g., *"ready to"* $\not\rightarrow$ *NULL* in Fig. (2i);

- hypothetical conjunctions, e.g., *"if"* $\not\rightarrow$ *NULL* in Fig. (2j).

The hypothesis novelty refers to the expression in H that conveys novel information against P. It is also marked through connecting it to *NULL* through a negative link. Such an expression is usually among the following cases:

- numbers, e.g., *NULL* $\not\rightarrow$ *"20-30 percent"* in Fig. (2k);

- novel content words, e.g., *NULL* $\not\rightarrow$ *"property damage"* in Fig. (2l).

4 Alignment-based RTE Methods

In this section, the conventional alignment-based RTE method is introduced first. This method is then augmented to leverage the proposed labeled alignment to improve the prediction accuracy.

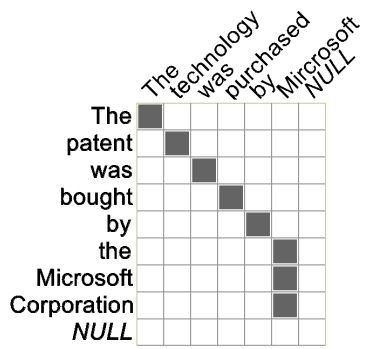### 4.1 RTE Method Based on Normal Alignment

The conventional alignment-based RTE method measures the quality of the alignment between the premise P and the hypothesis H to predict their entailment relation (Fig. 3a). An automated aligner is first learned from the annotation of positive links, the normal alignment consists of positive links (see Sec. 3.1). Then this aligner produces an alignment for each input (P, H). After that, a feature extractor measures the quality of the alignment. Finally a classifier utilizes these measures as features to predict the entailment relation. Commonly used quality measurements for alignment include the confidence score of the aligner and the ratio of linked words in P (Tab. 1).

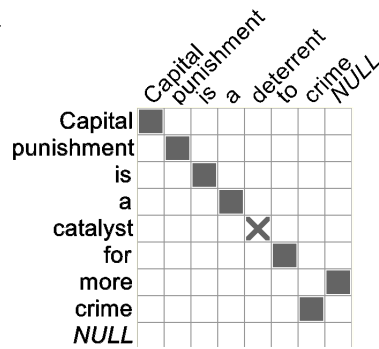### 4.2 RTE Method Based on Labeled Alignment

The augmented RTE method based on the labeled alignment not only measures the quality of the alignment, but also detects the signals of negative links to improve the prediction accuracy (Fig. 3b). The augmentation is conducted in two aspects. First, the aligner is trained with both positive and negative links, thus the produced alignment for each input (P, H) contains both positive and potentially negative links (but two types of links are not distinguished). Second, the feature extractor not only measures the quality of the alignment, but also analyzes the type of each link. A wide range of type-related features can be extracted from each link of the alignment (Tab. 1). These type-related features together with the quality-related features are added into a feature vector for classification.

Notably, besides the above RTE method, a pipeline framework based on the labeled alignment has been tried, but its accuracy turns to be lower than that of the baseline. The
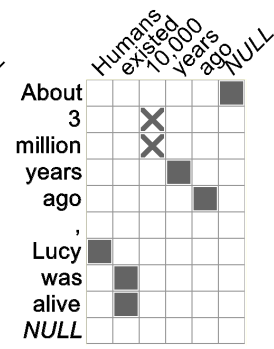
---

[4]This phenomenon is actually hard to recognize in a practical system. Multiple relevant weak features for classification are employed.
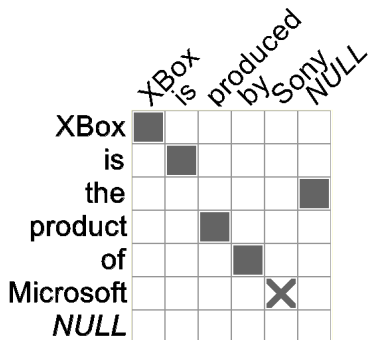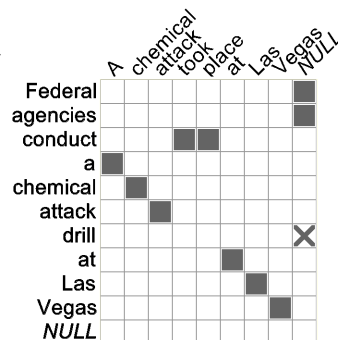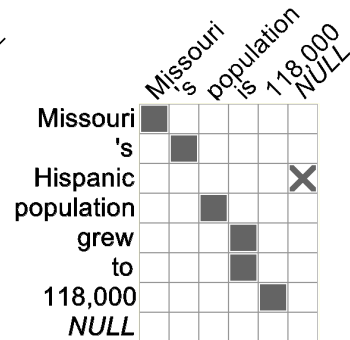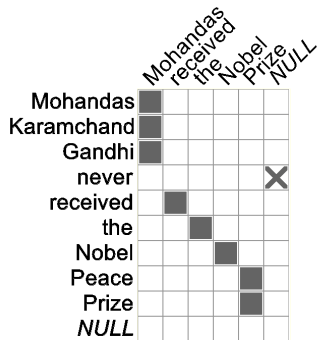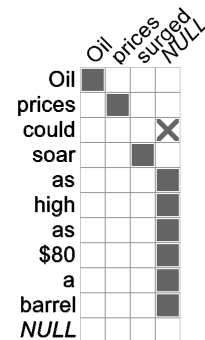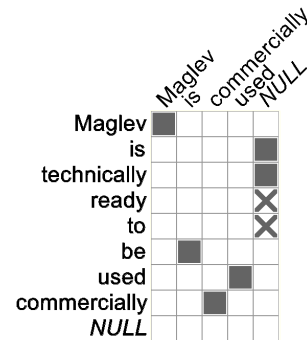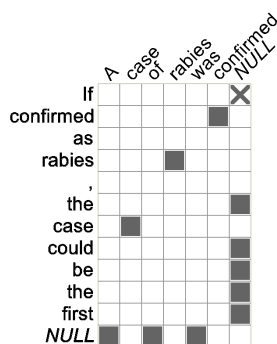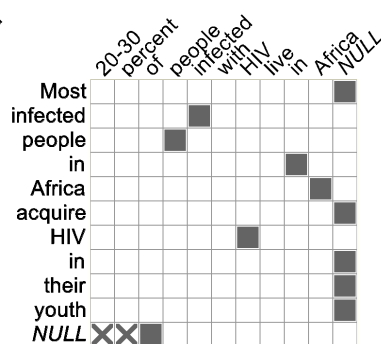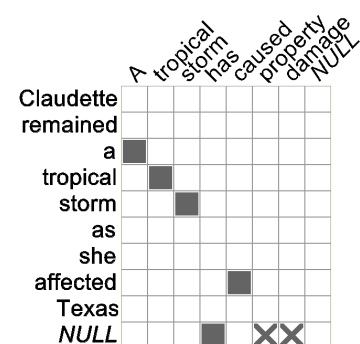
Figure 2: Examples of labeled alignment. Each subfigure presents an RTE sample. The vertical text is the premise, and the horizontal text is the hypothesis. The solid squares represent positive links, and the crosses represent negative links. (a) is of entailment relation, and (b)–(l) are of non-entailment relation.

(a) Baseline Method Based on Normal Alignment

(b) Proposed Method Based on Labeled Alignment Method
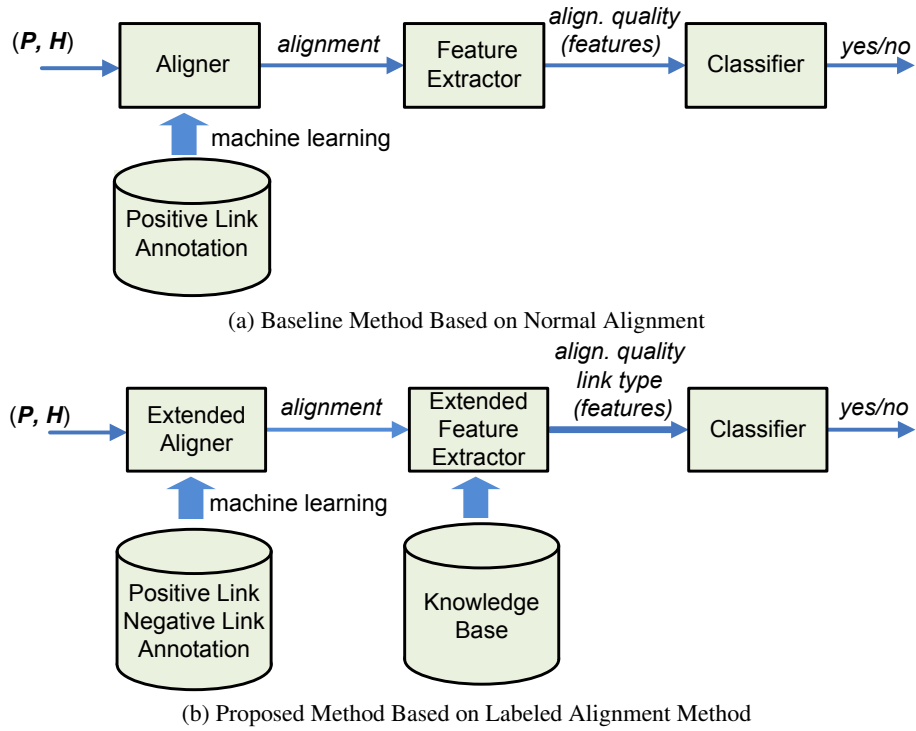
Figure 3: Baseline and Proposed Alignment-based RTE methods

Category	Feature
Align.	Confidence score of the aligner
Quality	Ratio of linked words in P
Link Type	Whether e_P and e_H are in an antonym list [a]
	Whether e_P and e_H are in an synonym list
	Whether e_P and e_H are unequal numbers
	Whether e_P and e_H are different named entities
	Relation of e_P and e_H in an ontology (hyponym, sibling, etc.)
	Ontology-based similarities of e_H and e_P
	Count of common characters
	Length of the common prefixes
	Length of the common suffix
	Tuple of the Part-of-Speeches [b]
	Tuple of the ancestors in an ontology
	Tuple of whether e_H or e_P is in a list of negative expressions
	Tuple of whether e_H or e_P is the head of a noun phrase

[a] Suppose the link is from e_P to e_H where e_P and e_H are the expressions in the premise P and the hypothesis H, respectively.

[b] Tuple features are the tuples of the values extracted from e_P and e_H, respectively.

Table 1: Features Extracted from Alignments for RTE Classification

	# Train.	# Test.	Ratio Posi.
RITE1	407	407	0.649
RITE2	814	781	0.596

Table 2: Experimental Data Sets

pipeline method first employs a classifier to predict whether each link is positive or negative, and then employs another classifier to predict the entailment relation based on the confidence scores of the first classifier.

5 Experiment

The data sets of the NTCIR-9[5] RITE1[6] and NTCIR-10 RITE2 shared tasks (simplified Chinese binary-class track) are taken as the experimental data sets(Shima et al., 2011; Watanabe et al., 2013). This section first describes the annotating process of the labeled alignment, then presents the experimental settings and finally reports the experimental results.

5.1 Data Set Annotation

The data sets from the simplified Chinese binary-class tracks of NTCIR-9 RITE1 and NTCIR-10 RITE2 contains 1,595 sentence pairs in all (Tab. 2). Note that all the training and test samples of RITE1 are reused as the training samples of RITE2, while newly collected 781 sentence pairs are taken as the test samples.

The annotating process follows the methodology employed by (Brockett, 2007). The training set of NTCIR-9 RITE1 is used for training annotators, and three Chinese native-speaking undergraduates are actively encouraged to discuss the arising cases, resolve questions and reconcile results with the authors. In annotating the test set of NTCIR-9 RITE1, however, they are first instructed not to discuss the annotations either with the authors or among themselves in order to measure annotator agreement. After that, they reconcile the results on the test set with the authors.

The measure of annotator agreement indicates the alignment annotations are reliably consistent. All three annotators concurred on about 72% of proposed links on the test set, two out of three

agreed on about 24% of cases, and three-way disagreements were as rare as about 4%.

5.2 Experimental Settings

The supervised learning aligner described in (Chambers et al., 2007) and (MacCartney et al., 2008) is adopted in this paper. This aligner is a structured learning algorithm that employs a linear weighted scoring function to evaluate each candidate alignment. We adapt the original algorithm from two aspects. First, the candidate alignment links are generated from a wide range of NLP analysis results, as follows,

- each segmented word in $P \rightarrow$ each segmented word in H;

- each syntactic node in $P \rightarrow$ each syntactic node in H;

- each NE in $P \rightarrow$ each NE in H;

- each expression e_P in $P \rightarrow$ each expression e_H in H as long as (e_P, e_H) appears in a synonym list, a antonym list, or an ontology.

Second, the alignment-learning features contains all the link type features in Tab. 1. These two enhancements, abstractly, convert aligning to a comprehensive NLP process.

The BaseSeg toolkit based on the conditional random field is employed to segment the Chinese texts (Zhao et al., 2006). The Stanford factored parser, which is reported to be more accurate than the PCFG parsers (Klein and Manning, 2002), is employed to analyze the segmented Chinese text. The BaseNER toolkit is employed to recognize named entities (Zhao and Kit, 2008).

We take two Chinese ontologies – CiLin[7] (Mei et al., 1983) and HowNet (Dong and Dong, 2003) – as the knowledge-base for extracting features. Three methods of computing the semantic similarity proposed in (Liu and Li, 2002; Xia, 2007) are employed.

We take the RBF-kernelled SVM as the entailment classifier. The implementation of LibSVM is employed. The parameters are tuned through 5-fold cross-validation on the training set.

The conventional RTE method based on the normal alignment, which is presented in Sec. 4.1, is taken as the baseline method.

---

[5] NTCIR is the abbreviation of NII Test Collection for IR Systems where NII abbreviates the National Institute of Informatics in Japan.

[6] RITE is the abbreviation of Recognizing Inference in TExt.

[7] This term means a word forest of synonyms in Chinese.

Method	Acc. on RITE1	Acc. on RITE2
Top entries	0.7764 (ICRC_HITSZ[a] Run03[b])	0.6850 (MIG Run02)[c]
	0.7617 (FudanNLP Run02)	0.6812 (CYUT Run03)
	0.7568 (ICRC_HITSZ Run02)	0.6658 (WHUTE Run02)
	0.7469 (FudanNLP Run01)	0.6581 (MIG Run01)
	0.7371 (WHUTE Run03)	0.6479 (WHUTE Run01)
normal align. (baseline)	0.7715	0.6991
labeled align. (proposed)	**0.8129**	**0.7465**

[a] Team ID;
[b] Run ID. Each team can submit the results of five runs at most.
[c] The top entry is the proposed method, thus not listed.

Table 3: Entailment Prediction Accuracy on NTCIR-9 RITE and NTCIR-10 RITE2 Data Sets

5.3 Experimental Results

The experimental results of the prediction accuracy on NTCIR-9 RITE1 and NTCIR-10 RITE2 data sets are presented at Tab. 3. The participants mainly employ committees of classifiers to learn from a wide range of features including multi-level similarities, occurrences of negative words, mismatches of named entities and numbers, syntactic correspondences, and so on (Zhang et al., 2011; Ren et al., 2011). The results show that the proposed RTE method outperforms not only the baseline method, but also the official entries of the shared tasks.

6 Conclusion and Future Work

In this paper, a labeled alignment scheme is proposed to address the shortage of the normal alignment scheme for non-entailment RTE samples. To verify the proposed scheme, an augmented alignment-based RTE method that employs the labeled alignment is compared with a conventional one that employs the normal alignment. The data sets of two shared RTE tasks are taken as the experimental data sets and manually annotated with the proposed scheme. Experimental results indicate that the augmented RTE method outperforms not only the baseline method, but also all the submitted systems of the shared tasks. Therefore, the proposed labeled alignment scheme proves to be effective.

The future work of this paper is two-fold. First, during the research, though two Chinese ontology resources – CiLin and HowNet – are employed to detect negative links, it is found that quite a few critical semantic relations are not covered. Therefore we plan to merge and scale existing Chinese ontologies through data mining techniques

such as (Liu and Singh, 2004). Second, the proposed method is actually applicable to multiple languages, though it is only tested on Chinese in this paper. We plan to apply it to other languages such as the Microsoft English RTE corpus in the future.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2009. A survey of paraphrasing and textual entailment methods. http://arxiv.org/abs/0912.3747. [accessed 10-Jan-2013].

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of HLT-EMNLP*, pages 628–635.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. *Microsoft Research Technical Report MSR-TR-2007-77*.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Zhen Dong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *Proceedings of International Conference on Natural*

Language Processing and Knowledge Engineering, pages 820–824. IEEE.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with LCC's GROUNDHOG system. In *Proceedings of the Workshop on the Second PASCAL Recognising Textual Entailment Challenge*.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15(2003):3–10.

Milen Kouylekov, Alessio Bosca, and Luca Dini. 2011. EDITS 3.0 at RTE-7. *Proceedings of the Seventh PASCAL Recognizing Textual Entailment Challenge*.

Qun Liu and Su Jian Li. 2002. Computation of semantical similarity for phrases based on HowNet (in Chinese). *Chinese Computational Linguistics*, 7(2):59–76.

Hugo Liu and Push Singh. 2004. Conceptnet – a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of HLT-NAACL*, pages 41–48.

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP'08*, pages 802–811.

Erwin Marsi and Emiel Krahmer. 2005. Classification of semantic relations by humans and machines. In *Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 1–6.

Yashar Mehdad. 2009. Automatic cost estimation for tree edit distance using particle swarm optimization. In *Proceedings of ACL-IJCNLP*, pages 289–292.

Jia Ju Mei, Yi Ming Zhu, and Yun Qi Gao. 1983. *TongYiCi CiLin*. Shanghai Dictionary Publisher.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastia Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of ACL-IJCNLP*, pages 297–305.

Han Ren, Chen Lv, and Donghong Ji. 2011. The WHUTE system in NTCIR-9 RITE task. In *Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan*.

Miguel Rios and Alexander Gelbukh. 2012. Recognizing textual entailment with a semantic edit distance metric. In *11th Mexican International Conference on Artificial Intelligence,*, pages 15–20. IEEE.

Mark Sammons, VG Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of ACL'10*, pages 1199–1208.

Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan*.

Alexander Volokh and Günter Neumann. 2011. Using MT-based metrics for RTE. In *Proceedings of the Seventh PASCAL Recognizing Textual Entailment Challenge*.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, pages 784–792. ACL.

Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa, Keita Nabeshima, Naoaki Okazaki, and Kentaro Inui. 2012. Leveraging diverse lexical resources for textual entailment recognition. *ACM Transactions on Asian Language Information Processing*, 11(4):18.

Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, CHeng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. 2013. Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop. In *Proceedings of NTCIR-10*.

Tian Xia. 2007. Research on the computation of semantical similarity for Chinese phrases (in Chinese). *Computer Engineering*, 33(6):191–194.

Yaoyun Zhang, Jun Xu, Chenlong Liu, Xiaolong Wang, Ruifeng Xu, Qingcai Chen, Xuan Wang, Yongshuai Hou, and Buzhou Tang. 2011. ICRC_HITSZ at RITE: Leveraging multiple classifiers voting for textual entailment recognition. In *Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan*.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165. Sydney: July.