BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RITE-2 Task*

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

ABSTRACT

In this paper, we propose a labeled-alignment-based RTE method to approach the simplified Chinese textual entailment track in the NTCIR-10 RITE-2 task. The labeled alignment, compared with the normal alignment, employs negative links to explicitly mark the contradictory expressions between the two sentences to justify the non-entailment pairs. Therefore, the corresponding alignment-based RTE method can gain accuracy improvement through actively detecting the signals of non-entailment. The performance of the proposed method in the formal run achieves Macro-F1's of 73.84%, 56.82%, and Worse Ranking (R) of 8.00% for the simplified Chinese subtasks of Binary Class (BC), Multi-Class (MC) and RITE for Question Answering (RITE4QA), respectively.

Team Name

bcNLP

Subtasks

RITE BC, MC and RITE4QA (Simplified Chinese)

Keywords

alignment, RTE, SVM

1. INTRODUCTION

The bcNLP team from the Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University, participated in the Binary Class (BC), Multi-Class (MC) and RITE for Question Answering (RITE4QA) subtasks of the NTCIR-10 Recognizing Inference in TExt (RITE) [11]. This paper describes our method and discusses the official results.

The principle of the alignment-based RTE methods is that a sufficiently good alignment between the premise t_1 and the hypothesis t_2 means a close lexical and structural correspondence, thus an entailment relation might exist between them. For example, Fig. (1a) shows that the entailment relation is correctly predicted through recognizing '*read into*' → '*interpreted*'[1] and '*what he wanted*' → '*in his own way*'.

However, the alignment scheme is mainly developed in MT, which does not solve the non-alignment samples well. It usually links the words in t_2, which have no counterparts in t_1, to $NULL$ regardless their impacts on the entailment relation. For example, in Fig. (1b), '*ferry sinking*', '*cause*' and '*that*' are all linked to $NULL$[2], while only '*ferry sinking*' is the cause for non-entailment. Therefore, such an alignment is insufficient for RTE.

This paper extends the alignment scheme to meet the challenge of RTE. The extended scheme, named labeled alignment, introduces another type of links, named negative links, to mark those critical linguistic phenomena that usually cause non-entailment relations. For example, Fig. (1c) shows that the previous vital expressions '*ferry sinking*' is linked to '*flood*' through a negative link, noted as '*ferry sinking*' $\not\to$ '*flood*'.

The rest of this paper is organized as follows. The proposed RTE method is presented in Sec. 2. Then the system is described in Sec. 3. The official evaluation is presented and discussed in Sec. 4. Finally Sec. 5 concludes this paper with a description of future work.

2. METHODS

In this section, the conventional alignment-based RTE method is introduced first. Then this method is extended to leverage the labeled alignment to improve the prediction accuracy.

2.1 RTE Method Based on Normal Alignment

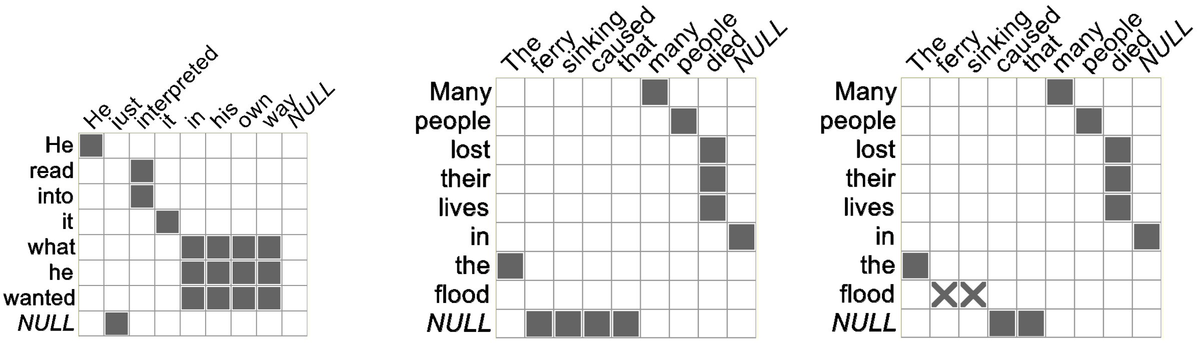
The conventional alignment-based RTE method measures the quality of the alignment between the premise t_1 and the hypothesis t_2 to predict their entailment relation (Fig. 2a). Its working flow is as follows.

- First, An automated aligner is learned from the annotation of normal alignment links.
- Then this aligner produces an alignment for each input (t_1, t_2).
- After that, a feature extractor measures the quality of the alignment.
- Finally a classifier utilizes these measures as the features to predict the entailment relation.

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 61170114, and Grant No. 61272248), and the National Basic Research Program of China (Grant No. 2009CB320901 and Grant No.2013CB329401).

[1] The notation means that the expression '*read into*' in t_1 is connected to the expression '*interpreted*' in t_2.

[2] $NULL$ means an empty expression.



(a) Alignment on entailment pair (b) Weakness on non-entailment pair (c) Labeled alignment on non-entailment pair

Figure 1: Illustration of Alignment for RTE. Each subfigure presents an RTE sample. The vertical text is the t_1, and the horizontal text is the hypothesis. The solid squares represent normal links, and the crosses represent negative links. (a) is of entailment relation, while (b) and (c) are of non-entailment relations.

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

Table 2: Experimental Data Sets

The frequently employed quality measurements for alignment include the confidence score of the aligner and the ratio of linked words in t_1 (Tab. 1).

2.2 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. 2b). The augmentation is conducted in two aspects. First, the aligner is trained with both the normal and the negative links, thus the produced alignment for each input (t_1, t_2) contains both positive and 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.

3. SYSTEM DESCRIPTION

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. We manually annotate the training set of RITE2 to train the automated aligner.

The supervised learning aligner described in [1] and [9] is adopted in this paper. This aligner employs a linear weighted scoring function to evaluate each candidate alignment, and a simulated annealing algorithm is employed to find the best alignment.

The BaseSeg toolkit based on the conditional random field is employed to segment the Chinese texts [13]. The Stanford factored parser, which is reported to be more accurate than the PCFG parsers, is employed to analyze the segmented Chinese text [4, 5]. The BaseNER toolkit is employed to recognize named entities [14].

Two Chinese ontologies – CiLin[3] [10, 8] and HowNet [3] – are taken as the knowledge-base for extracting features. Three methods of computing the semantic similarity proposed in [15, 7, 12] are employed.

The RBF-kernelled SVM is taken as the entailment classifier. The implementation of LibSVM is employed [2]. The parameters are tuned through 5-fold cross-validation on the training set.

The classification framework of the MC subtask is different from those of BC and RITE4QA. BC and RITE4QA are both binary problems, thus a single classifier is sufficient. MC is a multi-class problem, where the flat one-vs-rest framework is employed.

Three runs are submitted for each subtask of BC, MC and RITE4QA. These runs on different subtasks all adopt the same settings as follows. Run01 employs character overlap, n-gram (n=2,3) overlap as features, and employs an RBF-kernelled SVM to learn from these features to predict the entailment relation. Run02 is the conventional alignment-based RITE method described in Sec. 2.1. Run03 is the proposed labeled-alignment-based method described in Sec. 2.2.

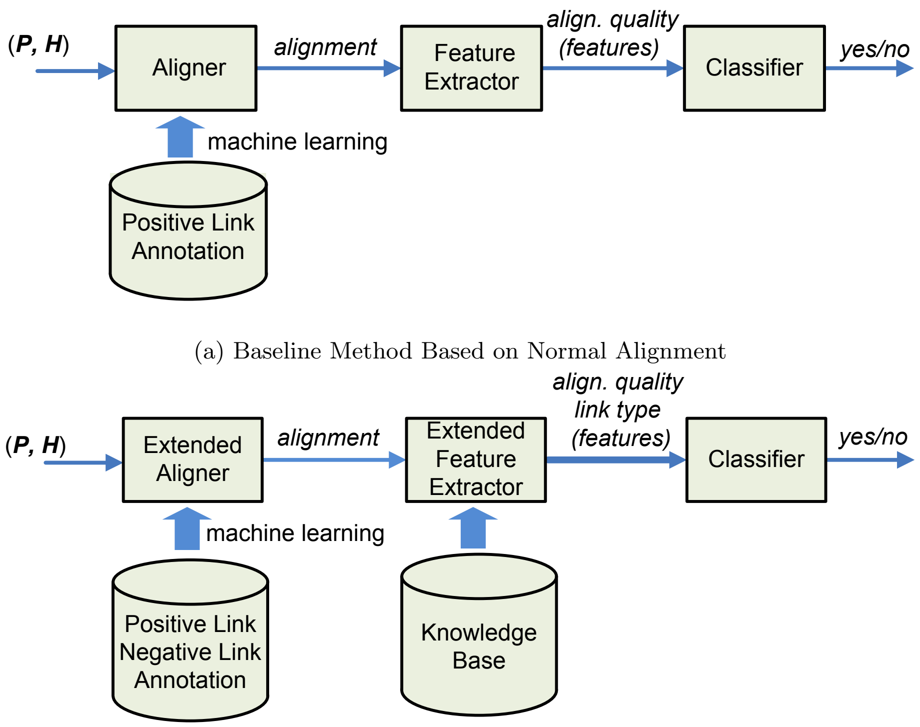
4. EVALUATION AND DISCUSSION

The official evaluation are presented at Tab. 3. The results show that the proposed RTE method (Run03) are quite effective on the subtasks of BC and MC. It not only significantly outperforms the baseline methods (Run01 and Run02), but also be competitive among the participants.

However, the performance of all our runs on RITE4QA is poor. We study the samples of RITE4QA, and notice that these samples are quite different from those of BC and MC [4]. The t_1's of the RITE4QA samples are usually much longer

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

[4] We focused on the BC and MC tasks when we were developing the RTE method. We just tried RITE4QA, and did not adapt the system due to lack of time



(a) Baseline Method Based on Normal Alignment

(b) Proposed Method Based on Labeled Alignment Method

Figure 2: Baseline and Proposed Alignment-based RTE methods

Category	Feature
Align. Quality	Confidence score of the aligner
	Ratio of linked words in t_1
Link Type	Whether e_1 and e_2 are in an antonym list [a]
	Whether e_1 and e_2 are in an synonym list
	Whether e_1 and e_2 are unequal numbers
	Whether e_1 and e_2 are different named entities
	Relation of e_1 and e_2 in an ontology (hyponym, sibling, etc.)
	Ontology-based similarities of e_1 and e_2
	Count of common characters
	Length of the common prefixes
	Length of the common suffix
	Tuple [b]of the syntactic tags [c]
	Tuple of the ancestors in an ontology
	Tuple of whether e_1 or e_2 is in a list of negative expressions
	Tuple of whether e_1 or e_2 is the head of a noun phrase

[a] Suppose the link is from e_1 to e_2, where e_1 and e_2 are the expressions in the premise t_1 and the hypothesis t_2, respectively.

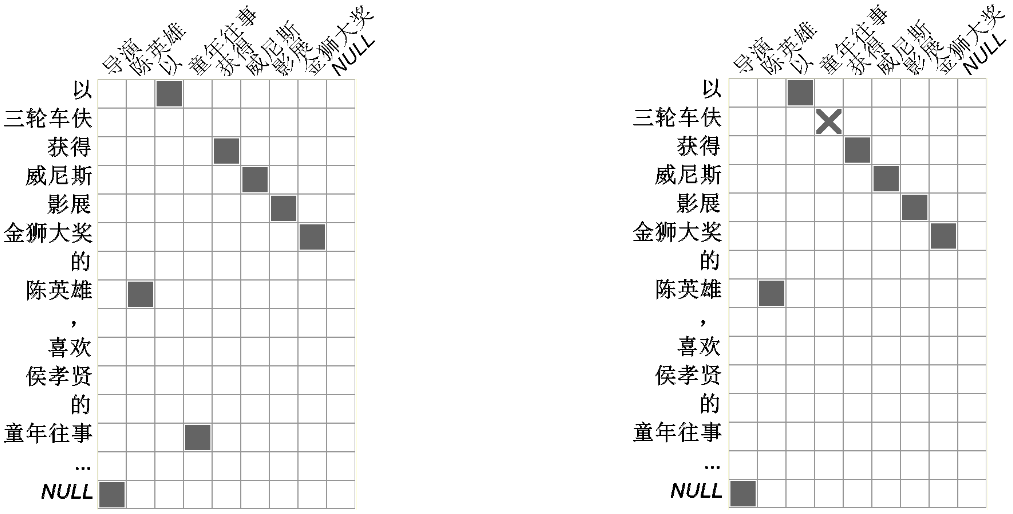
[b] Tuple features are the tuples of the values extracted from e_1 and e_2, respectively.

[c] We search for the node corresponding to the expression in the constituent parse tree. If such a node can not be found, the tag will be *NULL*.

Table 1: Features Extracted from Alignments for RTE Classification

Run	Macro-F_1 on BC	Macro-F_1 on MC	Worse Rank.(R) on RITE4QA
Run01	67.04	39.95	2.67
Run02	66.89	44.88	0.00
Run03	73.84	56.82	8.00

Table 3: Official Evaluation



(a) Fault Alignment on an RITE4QA sample

(b) Successful Alignment on the RITE4QA sample

Figure 3: Examples of Alignment for RITE4QA Samples. Each subfigure presents an RTE sample. The vertical text is t_1, and the horizontal text is t_2. The solid squares represent normal links, and the crosses represent negative links.

that those of the BC and MC samples. In addition, the t_1's of the RITE4QA samples usually contains all the words of the t_2's. However, the current manual alignments on the BC samples and the trained aligner give high priority on identical words. For example, Fig. 3a present the alignment produced our automated aligner on an RITE4QA sample, which leads to a false positive prediction. Fig. 3a presents the ideal alignment that we would annotate to train the aligner.

5. CONCLUSIONS

In this paper, a labeled alignment scheme is proposed to address the shortage of the normal alignment scheme for non-entailment RTE samples. The official evaluation indicates that the augmented RTE method achieves high accuracy on the BC and MC subtasks, while is not successful on the RITE4QA subtask.

Our future work is two-fold. First, the application of the proposed method is not successful on the RITE4QA subtask. We plan to annotate the data set of the RITE4QA samples to re-train the automated aligner, with which higher prediction accuracy might be achieved. Second, 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 have not been covered for the task. Therefore we plan to merge and scale up existing Chinese ontologies through data mining techniques such as [6].

6. REFERENCES

[1] N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, M.-C. de Marneffe, D. Ramage, E. Yeh, and C. D. Manning. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170. Association for Computational Linguistics, 2007.

[2] C. C. Chang and C. J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[3] Z. D. Dong and Q. Dong. Hownet-a hybrid language and knowledge resource. In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, pages 820–824. IEEE, 2003.

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

[5] R. Levy and C. Manning. Is it harder to parse Chinese, or the Chinese Treebank. In *Proceedings of ACL*, volume 3, pages 439–446, 2003.

[6] H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[7] Q. Liu and S. J. Li. Computation of semantical

similarity for phrases based on HowNet (in Chinese). *Chinese Computational Linguistics*, 7(2):59–76, 2002.

[8] Z. C. Luo. Improvements on TongYiCi CiLin. http://blog.csdn.net/ganlantree/article/details/1845788, 2007. [accessed 10-Jan-2013].

[9] B. MacCartney, M. Galley, and C. D. Manning. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 802–811. Association for Computational Linguistics, 2008.

[10] J. J. Mei, Y. M. Zhu, and Y. Q. Gao. *TongYiCi CiLin*. Shanghai Dictionary Publisher, 1983.

[11] 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.

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

[13] H. Zhao, C. N. Huang, and M. Li. 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, 2006.

[14] H. Zhao and C. Y. Kit. 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, 2008.

[15] Y. H. Zhu, H. Q. Hou, and Y. T. Sha. Comparison and evaluation of two algorithms for recognizing Chinese synonyms (in Chinese) . *Journal of Library Science in China*, 28(4):82–85, 2002.