HITSZ-ICRC: An Integration Approach for QA TempEval Challenge

Yongshuai Hou, Cong Tan, Qingcai Chen and Xiaolong Wang
Department of Computer Science and Technology
Harbin Institute of Technology Shenzhen Graduate School
HIT Campus, The University Town of Shenzhen, Shenzhen, 518055, China
{yongshuai.hou, viptancong, qingcai.chen}@gmail.com
wangxl@insun.hit.edu.cn

Abstract

This paper presents the HITSZ-ICRC system designed for the QA TempEval challenge in SemEval-2015. The system used an integration approach to annotate temporal information by merging temporal annotation results from different temporal annotators. TIPSemB, ClearTK and TARSQI were used as temporal annotators to get candidate temporal annotation results. Evaluation demonstrated that our system was effective for improving the performance of temporal information annotation, and achieved recalls of 0.18, 0.26 and 0.19 on Blog, News and Wikipedeia test sets.

1 Introduction

The QA TempEval (Llorens et al., 2015) in SemEval-2015 is a temporal information annotation challenge, which is a follow-up task after TempEval-1 (Verhagen et al., 2007), TempEval-2 (Verhagen et al., 2010) and TempEval-3(UzZaman et al., 2013). QA TempEval task is similar to the task ABC in TempEval-3, requires participant system (1) extracting and normalizing temporal expressions, (2) extracting events and (3) identifying temporal relations on plain documents. Temporal information annotation should follow TimeML scheme (Pustejovsky et al., 2003a). Difference between QA TempEval task and task ABC in TempEval-3 is evaluation method: in all previous TempEval tasks, annotated result was evaluated by the temporal information annotation accuracy based on manually annotated test corpus; in QA TempEval, annotated result was evaluated by temporal question-answering(QA) accuracy in the given temporal QA system (UzZaman et al., 2012) based on temporal knowledge produced from participant's annotation.

Temporal annotation is useful in information retrieval, QA, natural language understanding and so on. A lot of researches have been attracted on this topic in the past years. Many methods were proposed and many toolkits were implemented for temporal information annotation.

TIMEN (Llorens et al., 2012a) is a community-driven tool using rule-based method based on knowledge base to solve the temporal expression normalization problem. TARSQI Toolkit (Verhagen and Pustejovsky, 2008) is a modular system for automatic temporal information annotation. The toolkit can extract temporal expressions, events and recognize temporal relations by its different components. Llorens et al. (2010) used CRF models based on semantic information to annotate temporal information according to TimeML scheme, and their TIPSem system got outstanding performance results in TempEval-2. Steve (2013) piped machine-learning models in his ClearTK system to annotate temporal information using a small set of features. His system got best performance for temporal relation identification in TempEval-3. The TIMEN toolkit was integrated into the ClearTK system for temporal expression normalization. Llorens et al. (2012b) proposed an automatic method to improve the correctness of each individual annotation by merging different annotation results with different strategies.

This paper described the method HITSZ-ICRC system used for QA TempEval challenge. This was first time for HITSZ-ICRC team to do the temporal annotation task. An integration approach was chosen to get improved annotation result on currently available temporal annotation toolkits for QA TempEval task. Annotation results from those

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 830–834,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

toolkits were merged using a temporal annotation merging method (Llorens et al., 2012b).

The remainder of this paper is structured as follows: Section 2 describes the system modules used for temporal information annotation. Section 3 introduces the data sets and toolkits used, explains and analysis the evaluation results. Section 4 concludes the paper.

2 Integration Approach for Temporal Information Annotation

QA TempEval task required participant system to annotate temporal expressions, events and temporal relations following TimeML scheme.

Many toolkits are available for temporal information annotation, such as TARSQI (Verhagen and Pustejovsky, 2008), ClearTK (Bethard, 2013) TIPSemB (Llorens et al., 2010) and so on. Each toolkit can be used as a temporal annotator to get candidate annotation result.

But annotation results from current toolkits cannot be used for QA TempEval directly because some annotations do not in the TimeML format. For example, time expression normalization values in some results are in independent format, such as "*20140804AF*", should be as "2014-08-04TAF"; some time expressions are not normalized and are set to "*default_norm*" or no value; some toolkits change source text content after annotating temporal information, such as changing adjacent spaces to single space. So an annotation corrector module is necessary to correct candidate annotation results.

Automatic method proposed by Llorens et al. (2012b) was employed to merge annotation results from different annotators. The method used weighted voting techniques to merge temporal annotations. Weight for each candidate result and threshold for choosing final annotation were variable. Element in merged result should get weight above the threshold. Based on different weight and threshold settings, merged results can satisfy different requirements: such as high recall, high precision and balanced precision and recall.

Annotation toolkits and the results merging method were used to get final annotation result. Steps to get final result are as follows:

Step1: re-training models with train dataset for temporal annotator;

Step2: annotating temporal information on test data using each annotator;

Step3: correcting annotation results from all annotators using temporal annotation corrector;

Step4: integrating all candidate annotation results to get final temporal annotation result using temporal result merger.

The temporal information annotation process of our system is shown in figure 1.

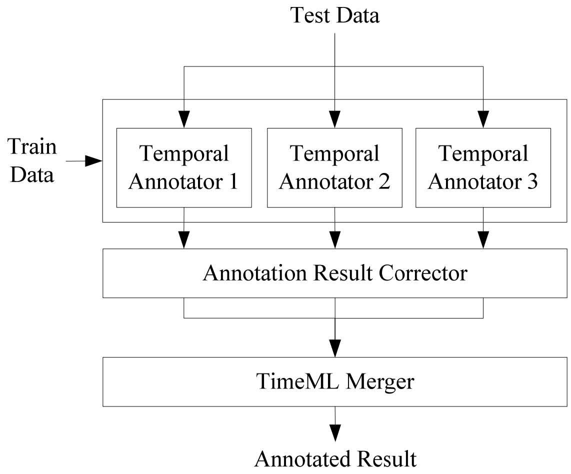

Figure 1. Temporal annotation process.

Annotation module used three temporal annotators here. The function of this module is for getting candidate temporal annotation results using different annotators.

Corrector module corrects all annotated results following TimeML scheme. Its functions include: (1) changing format of temporal expression values to TimeML format; (2) normalizing temporal expressions which have no value; (3) removing temporal expression tags which cannot be normalized, and removing the related temporal links at same time; (4) removing temporal entity tags with class labels not in TimeML label set and removing the related temporal links; (5) removing temporal links with class labels not in TimeML label set; (6) correcting the text content to source text.

The TimeML merger module used the temporal annotation merging method to merge annotation results. The F1 value for different annotators evaluated on develop data was used as voting weights. For QA TempEval task, high recall annotation result will be more effective, so high recall settings for the merging method were chosen. Different weight and threshold setting strategies were tried, which include: (1) Best F1 prior voting: the annotation chose as final result should be annotated by the best F1 annotator or at least two annotators; (2) Better F1 prior voting: the annotation chose as final result should be annotated by at least one annotator except the worst F1 annotator; (3) Union: the annotation chose as final result should be annotated by at least one annotator.

3 Results Evaluation

3.1 Dataset and toolkits

Train dataset provided for QA TempEval task is the same dataset in TempEval-3, includes TBAQ-cleaned dataset and TE3-Platinum (UzZaman et al., 2013) dataset. TBAQ-cleaned contains cleaned and improved AQUAINT and TimeBank corpus (Pustejovsky et al., 2003b). The TE3-Platinum is the evaluation corpus for TempEval-3 manually annotated by organizers. All the datasets are annotated in TimeML format.

The test dataset was in TempEval-3 format, and includes 28 plain text documents in Blog (8 documents), News (Wikinews, NYT, WSJ) (10 documents) and Wikipedia (10 documents).

Results evaluation was based on 294 temporal questions, 65 questions for Blog documents, 99 for News and 130 for Wikipedia. The question set was created by human experts based on the test documents. Annotated result was evaluated by the temporal QA system (UzZaman et al., 2012) using the question set.

The three annotation toolkits TARSQI, ClearTK and TIPSemB were used as temporal annotators. Default models in the toolkits were used for TARSQI and TIPSemB. Models in ClearTK were re-trained with the training data. In merging step, the temporal annotation merging toolkit (Llorens et al., 2012b) was used to get the final result.

3.2 Measures

Answers' precision (*P*), recall (*R*), and F1 value (*F1*) of the temporal QA system are used to evaluate annotation results. Recall is used as the main metric to sort results and F1 is used as secondary metric.

P, *R* and *F1* are calculated as:

$$P = \frac{num\_correct}{num\_answered} \quad (1)$$

$$R = \frac{num\_correct}{num\_questions} \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

where *num_correct* is the number of questions correctly answered by the temporal QA system based on temporal knowledge produced from participant's annotation result; *num_answered* is the number of questions answered by the temporal QA system based on participant's annotation result; *num_questions* is the number of test questions used in the temporal QA system.

3.3 Evaluation results with QA TempEval

Giving a test document, firstly it was annotated by three temporal annotators separately, including ClearTK, TIPSemB and TARSQI; then the annotated results were corrected to follow TimeML scheme by corrector module and were used as candidate results; finally, the three candidate results were merged using three different strategies. The models in ClearTK toolkit were trained with TBAQ-cleaned dataset.

Six results from different annotators and merging strategies were compared, including three results annotated by different annotators and three results annotated by different merging strategies. For the system had not been finished before submission deadline, only the result of TARSQI was submitted to QA TempEval challenge.

The evaluation results for the six temporal annotation results are shown in table 1, 2 and 3 in domain Blog, News and Wikipedia separately. *awd%* is the percentage of the answered questions and *corr* is the number of correct answers.

Run *TARSQI*, *TIPSemB* and *ClearTK* is the result annotated by corresponding temporal annotator. Run *BSTF_VOTE*, *BTRF_VOTE* and *RES_UNION* is the result produced with different merging strategies.

F1 value of each annotator result was used as its weight in merging step. *BSTF_VOTE* is the result merging with best F1 prior voting strategy. *BTRF_VOTE* is the result with better F1 prior voting strategy. R*ES_UNION* is the result with union strategy.

Results in table 1, 2 and 3 shows that performance of all merged results are better than results annotated by single annotator in each test domain. It means integration approach is effective for improving temporal information annotation performance. The union strategy performs best in all the six run results in all domains. So merging results from all annotators with union strategy is an effective way to get better annotation results based on QA TempEval evaluation method.

Run	Measures			Questions	
	P	R	F1	awd%	corr
TARSQI	0.17	0.02	0.03	0.09	1
TIPSemB	0.37	0.11	0.17	0.29	7
ClearTK	**0.55**	0.09	0.16	0.17	6
BSTF_VOTE	0.34	0.17	0.23	0.49	11
BTRF_VOTE	0.30	0.15	0.20	0.51	10
RES_UNION	0.36	**0.18**	**0.24**	**0.51**	**12**

Table 1. Evaluation results on Blog test data.

Run	Measures			Questions	
	P	R	F1	awd%	corr
TARSQI	0.47	0.08	0.14	0.17	8
TIPSemB	**0.55**	0.18	0.27	0.33	18
ClearTK	0.53	0.08	0.14	0.15	8
BSTF_VOTE	0.51	0.24	0.33	0.47	24
BTRF_VOTE	0.49	0.23	0.32	0.47	23
RES_UNION	0.51	**0.26**	**0.35**	**0.52**	**26**

Table 2. Evaluation results on News test data.

Run	Measures			Questions	
	P	R	F1	awd%	corr
TARSQI	**0.83**	0.08	0.14	0.09	10
TIPSemB	0.41	0.11	0.17	0.26	14
ClearTK	0.57	0.06	0.11	0.11	8
BSTF_VOTE	0.48	0.18	0.26	**0.37**	23
BTRF_VOTE	0.48	0.18	0.26	**0.37**	23
RES_UNION	0.54	**0.19**	**0.28**	0.35	**25**

Table 3. Evaluation results on Wikipedia test data.

Evaluation results show that annotation results from different annotators could be used to improve temporal information annotation performance by results merging. The precision of all merging results cannot achieve to the highest, and are lower than some annotator results. It means that the merging step merged wrong annotation into final result. The merging strategies tried in our experiments were more effective on improving the recall of temporal information annotation, which increased the chance that the temporal question could be answered, but were useless for question answering precision. So balancing the precision and recall is necessary for improving the performance of annotation results merging. Improving performance of single annotator also is important job for getting better final annotation result. We have tried the integration approach using results of the top 3 best performance systems in QA TempEval challenge(Llorens et al., 2015), and the result still can be improved.

4 Conclusions

We used an integration approach to annotate temporal information in HISZ-ICRC system for QA TempEval challenge. Annotation results from different annotators were merged using automatic merging method with different strategies. Evaluation results showed that the integration approach for temporal information annotation can effectively improve annotation performance than single annotator. Union strategy performed best in all strategies we tried.

We used same weight for temporal expression, event and temporal relation merging. But performance of different annotation modules is different in an annotator. We will try different weight setting for temporal expression, event and temporal relation annotation merging in future work. And the precision and recall have not been tried as merging weight in our experiment, which also will be tried in future work.

Acknowledgments

The authors thank Hector Llorens and Naushad UzZaman for sharing the TIPSemB and the temporal annotation merging toolkit, and thank the anonymous reviewers for their insightful comments.

This work was supported in part by the National Natural Science Foundation of China (61272383 and 61173075), the Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420 and JCYJ20120613151940045) and the Key Basic Research Foundation of Shenzhen (JC201005260118A).

References

Hector Llorens, Naushad UzZaman, and James Allen. 2012b. Merging Temporal Annotations. In *2012 19th International Symposium on Temporal Representation and Reasoning (TIME)*, pages 107–113, Leicester, UK, 12-14 September.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden, 15-16 July

Hector Llorens, Leon Derczynski, Robert J Gaizauskas, and Estela Saquete. 2012a. TIMEN: An Open Temporal Expression Normalisation Resource. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3044–3051, Istanbul, Turkey, May.

Hector Llorens, Nate Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 Task 5: QA TEMPEVAL - Evaluating Temporal Information Understanding with QA. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, USA.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and others. 2003b. The timebank corpus. In *Corpus linguistics*, 2003:40.

Marc Verhagen and James Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. In *22nd International Conference on Computational Linguistics: Demonstration Papers*, pages 189–192, Manchester, August.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, 15-16 July.

Naushad UzZaman, Hector Llorens, and James Allen. 2012. Evaluating Temporal Information Understanding with Temporal Question Answering. In *2012 IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 79–82.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, 14-15 June.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA, 14-15 June.