QA TempEval: Evaluating Temporal Information Understanding with QA

Hector Llorens, Nuance Communications, Alicante, Spain
Nate Chambers, United States Naval Academy , USA
Naushad UzZaman, Nuance Communications, NY, USA
Nasrin Mostafazadeh, University of Rochester, USA
James Allen, University of Rochester, USA
James Pustejovsky, Brandeis University, USA

Abstract

QA TempEval is a follow up of TempEval series in SemEval. It introduces a major shift in the evaluation methodology, which changes from temporal information extraction to temporal question-answering (QA). QA represents a natural way to evaluate temporal information understanding and creating tests requires much less expertise and effort. This allows us to perform, in addition to the traditional test created by the organizers, a crowd-sourced test created by the participants. Although the evaluation changes, the task for participating systems remains the same as previous editions, extracting temporal information from plain text documents (TempEval-3 format is re-used). The only difference is that this time systems' outputs are not compared to a human annotated key. Participant annotations are used to build a knowledge base for obtaining answers for temporal questions about the documents and these are compared to a human answer key. QA score measures performance in terms of the capacity of an approach to capture temporal information relevant to perform an end-user task, as compared with corpus-based evaluation where all information is equally important for scoring.

Introduction

QA TempEval is a follow up of TempEval series in SemEval: TempEval-1 [Verhagen et al., 2007], TempEval-2 [Verhagen et al., 2010], and TempEval-3 [UzZaman et al., 2013]. TempEval focus is on evaluating systems that extract temporal expressions (timexes), events and temporal relations as defined in TimeML standard [Pustejovsky et al., 2003] (timeml.org).

TimeML was developed to support research in complex temporal QA within the field of artificial intelligence (AI). Given the complexity of temporal QA, most of the efforts have focused, so far, on temporal IE, which has been evaluated with corpus-based evaluation in previous TempEvals. QA represents a natural way to evaluate temporal information understanding [UzZaman et al., 2012], and creating question sets is less complex for humans than manually annotating temporal information, which is required to perform corpus-based evaluation. Additionally, QA performance better captures the understanding of important temporal information as compared to corpus-based evaluation where all information is annotated and equally important for scoring.

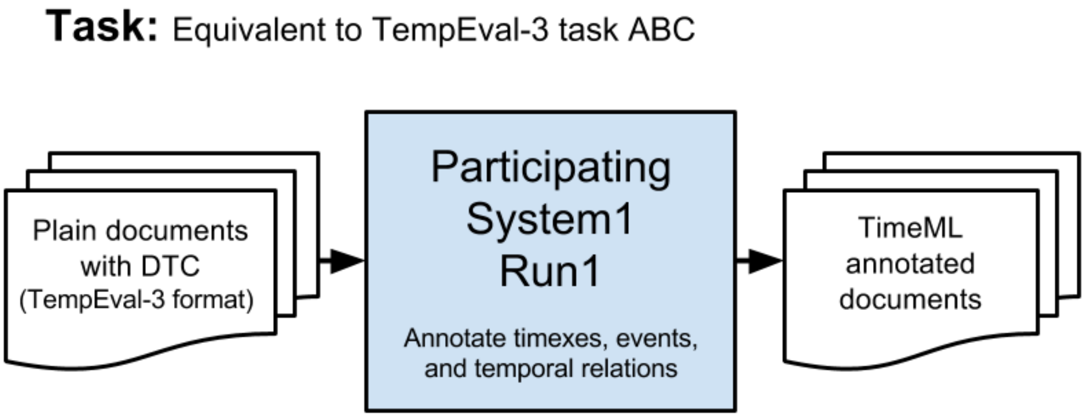
In QA annotators are not required to tag each individual event and try to order it regardless its importance but directly look for temporal relations that seem relevant or interesting in the document. Furthermore, in QA, annotators are not limited relations between entities in the same sentence or consecutive sentences. They can ask questions about any pair of entities in the document if a question comes naturally to a reader's mind, e.g., "did the election happen (e3) before the president gave (e27) the speech".

In QA TempEval, like in TempEval-3 task ABC, systems are required to perform end-to-end TimeML annotation from plain text and the complete set of temporal relations [Allen, 1983] is used.

Unlike TempEval-3, QA TempEval evaluation changes from IE to QA. Instead of IE performance a score based on QA performance is used to rank systems. Subtasks focusing on specific elements individually disappear in QA TempEval. Finally, in addition to news domain, Wikipedia articles and blog posts are included in the exercise.

Task Description

The task is equivalent to TempEval-3 task ABC, see Figure 1.


Task: Equivalent to TempEval-3 task ABC
Plain documents with DTC (TempEval-3 format)
Participating System1 Run1
Annotate timexes, events, and temporal relations
TimeML annotated documents

A set of text documents in TempEval-3 input format is given to participants. Participating systems are required to annotate the plain documents following TimeML scheme, which is divided in two types of elements:

- Temporal **entities**: These include **events** (EVENT tag, "came", "attack") and temporal expressions (**timexes**, TIMEX3 tag, e.g., "yesterday", "8 p.m.") as well as their attributes such as event classes and timex types and normalized values.
- Temporal **relations**: A temporal relation (**tlink**, TLINK tag) describes a pair of entities and the temporal relation between them. TimeML relations can be mapped to the 13 Allen interval relations as follows: SIMULTANEOUS and IDENTITY (equal), BEFORE (before), AFTER (after), IBEFORE (meets), IAFTER (meet-by), IS INCLUDED (during), INCLUDES (contains) and DURING (~), BEGINS (starts), BEGUN BY (started by), ENDS (finishes), ENDED BY (finished by), - (overlaps), - (overlapped by). For example, in (2), "6:00 pm" begins the state of being "in the gym".

  (2) John was in the gym between 6:00 p.m and 7:00 p.m.

  Note that TimeML does not explicitly include the Allen's overlap and overlapped by relations. However, these can be implicitly present in the temporal representation of a TimeML document by the combination of other relations. TimeML DURING has not clear mapping so we map it to SIMULTANEOUS (equal) for simplicity.

Each system's annotations represent its temporal knowledge of the documents. That knowledge is used as input for a temporal QA system [UzZaman et al. 2012] that will try to answer questions.

QA Evaluation Methodology

The main difference with previous TempEval editions is that the systems are not scored regarding how similar is their annotation to a human annotated key, but how useful is their TimeML annotation to answer human annotated temporal questions.

There are different kinds of temporal questions that could be answered given a TimeML annotation of a document. However, this first QA TempEval focuses on yes/no questions in the following format:

IS <entityA> <RELATION> <entityB> ?

(e.g., is event-A before event-B ?)

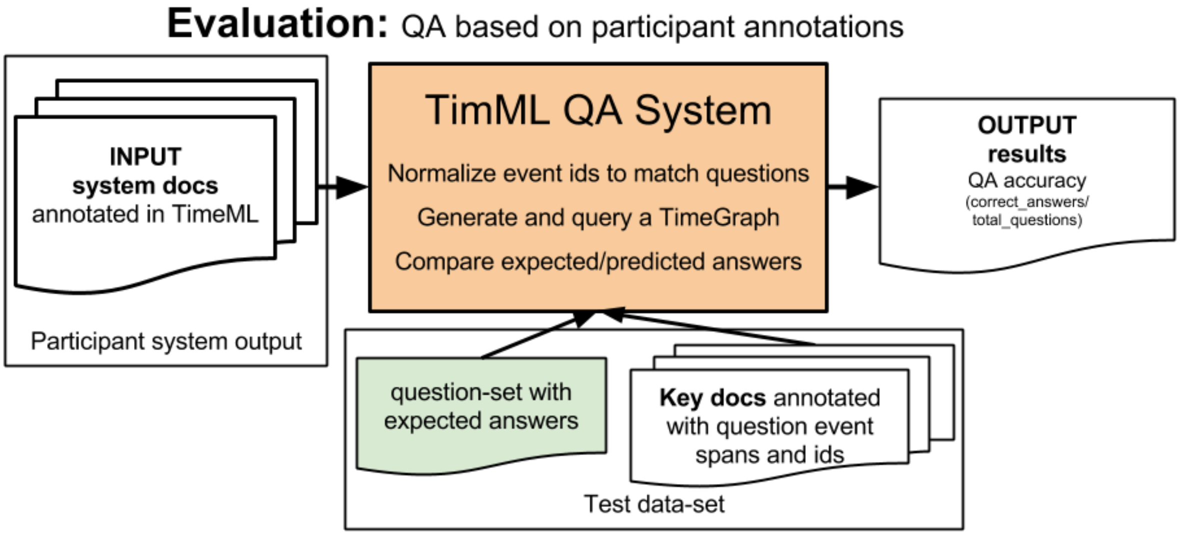
This makes it easier for human annotators to create accurate question sets with their answers. Other types of questions such as list-based make it more difficult and arguable in edge cases (e.g., list events between event-A and event-B). Questions about events not included in the document are not possible but we could ask about any time reference. However, this requires finding a close existing timex in the document and depending on the one we use to check the relation the answer could vary, so these are not included either.

The questions can be about any of these thirteen relations: BEFORE, AFTER, IBEFORE, IAFTER, BEGINS, BEGUN_BY, ENDS, ENDED_BY, OVERLAPS*, OVERLAPED_BY*, IS_INCLUDED, INCLUDES, SIMULTANEOUS.

Note: IDENTITY is equivalent to SIMULTANEOUS.

* Overlaps and Overlapped by cannot be explicitly annotated in TimeML but they could happen implicitly (i.e., be inferred from other relations).

The evaluation process is illustrated in Figure 2.



After the testing period, the participants send their TimeML annotations of the test documents. Organizers evaluate the TimeML annotations of all the participating systems with a set of questions. The systems are scored comparing the expected answers provided by human annotators against the predicted answers obtained from system's TimeML annotations.

Given system's TimeML annotated documents the process consists on three main steps:

- **ID Normalization**: Entities are referenced by TimeML tag ids, eids (e.g., eid23). The yes/no questions must contain two events with IDs (e.g., "*is event[eid23] after event[eid99] ?*"). The events of the question are annotated in a key document . However, systems may provide different ids to the same events. Therefore, the QA system normalizes the system annotated doc IDs with the question IDs that are annotated in the key docs using the TempEval-3 normalization tool, so that they match the key. The rest of the elements are also normalized accordingly.
- **Timegraph Generation**: The normalized TimeML docs are used to build a graph of time points representing the temporal relations of the events and timexes identified by the system. The timegraph is initialized by adding the TimeML explicit relations. With the timegraph's reasoning mechanism, the derived implicit relations are inferred. A very basic example follows. Given eventA is_before eventB and eventB is_before eventC, then the implicit relation eventA is_before eventC can be inferred. Timegraph

allows therefore answering questions about both explicit and implicit temporal relations. It can answer questions about any of the Allen relations.

- **Question Processing**: Answering questions requires temporal information understanding and reasoning. Note that asking: IS <entity1> <relation> <entity2> ? is not only asking if there is that explicit tlink between them but also, if it is not, if that relation can be inferred from other tlinks implicitly. Unlike corpus based evaluation, the system gets credit if its annotations provide the correct answer regardless of whether it annotates other irrelevant information or not. To answer the questions about TimeML entities (based on time intervals) using timegraph (based on time points), we convert the queries to point based queries. For answering yes/no questions, we check the necessary point relations in timegraph to verify an interval relation. For example, to answer the question is event1 after event2, our system verifies whether start(event1) > end(event2); if it is verified then the answer is true (YES), if it conflicts with the timegraph then it is false (NO), otherwise it is UNKNOWN.

QA Scoring

The possible human *answers* for a question are *yes*, *no*, or *unknown*. The later meaning that the answer can not be inferred from the document.

For each question we compare the obtained answer from the Timegraph (created with system annotations) and the expected answer (human annotated). The scoring is based on the following algorithm:

```
num_correct=0
num_answered=0

for question in question-set:
    num_questions += 1
    if (predicted_answer == yes or predicted_answer == no) or
    (annotated_answer == unknown):
        if predicted_answer == annotated_answer:
        # result = correct
            num_correct += 1
            num_answered += 1
        else:
        # result = incorrect
            num_answered += 1
    #else: predicted_answer == unknown and annotated_answer !=
    unknown
```

With this scoring system we calculate the following measures:

- Precision (P) = num_correct / answered_questions
- Recall (R) = num_correct / num_questions
- F measure (F1) = 2*P*R / P + R
  We will also pay attention to the percentage of questions answered, i.e., whose answer is different than unknown (Coverage = num_correct + num_incorrect / num_questions).

Datasets

Format

In QA TempEval, the creation of training/testing data does not require the manual annotation of all TimeML elements in source documents. After selecting the documents, for humans, it just takes reading the document, making temporal questions, providing the correct answers, and identifying the events included in the questions by bounding them in the text and adding an ID.

The format of the question sets is as follows:
<question-num>|<source_doc>|<question-with-ids>|<NL-question>|<answer>|[opt_extra_info]

An example would be:

3|APW19980418.0210.tml|IS ei21 AFTER ei19|Was Farkas cited by the police after becoming a brigadier general?|YES

APW19980418.0210.tml (KEY)
...Farkas <event eid="e19">became</event> a brigadier general
… He was <event eid="e21">cited</event> by the cops...

APW19980418.0210.tml (system annotation, full-TimeML)
...Farkas <event eid="e15" ...>became</event> a brigadier general
… He was <event eid="e24" ...>cited</event> by the cops...
<tlink eventID=e15 relatedToEventID=e24 relType=before /> ...

The system ids are normalized using the TimeML normalization tool used in TempEval-3. Then the answer obtained by the QA system is compared with the question-set answer.

Dataset contents

The main target language of the evaluation is English. As a novelty, QA tempeval includes wiki articles in addition to news domain. There are not much TimeML annotations outside the news domain but all the participants are in the same situation, so the evaluation would still be fair.

Training data consists of TempEval-3 TimeML annotated data. Furthermore, a question-set of 79 questions is provided so participants have data to test their systems. Furthermore, participants can easily extend the question-sets making questions over existing TimeML annotated data including TBAQ: TimeBank, AQUAINT (TempEval-3 training) and TE-3 Platinum (TempEval-3 testing).

- **Regular Training. 79 Yes/No questions and answers about the documents**
    - Revision of a dataset based on TimeBank [UzZaman et al. 2012]
    - Includes full TimeML annotation because source documents are from TimeBank
- **Social training.** Participants are encouraged to provide more questions and answers and share them as follows. A participant can select an existing TimeML annotated document (e.g., TimeBank document) and ask questions about the annotated events and provide the expected answer.
    - If the question can be answered given the TimeBank annotation then the question is valid.
    - Otherwise, the author can either improve TimeBank annotations or discard the question.

  The participant sends that data to the organizers and after a review process it will be added to the "social-training" dataset. This is a feasible task because reviewing questions and answers about documents is much simpler than reviewing complete TimeML annotations.

Regular Test corpora are based on the Newspapers (Wikinews, NYT, WSJ), Wikipedia and Informal blogs. Human experts are in charge of selecting the documents, creating the set of questions and answers about them, annotating the events in key documents, and finally reviewing the dataset correctness.

- **TE3-input text documents** (plain text and DCT): 30 documents selected by organizers
        - **10 News (Wikinews, WSJ, NYT)**: This covers the traditional TempEval domain used in all the previous editions.
        - **10 Wikipedia (history, biographical)**: This covers documents about people or history (en.wikipedia.org), which are rich in temporal entities.
        - **10 Blog (narrative)**: The blog dataset is coming from the Blog Authorship Corpus[1] [Schler et al.]. We hand selected blogs that are in narrative nature. For instance, blog entries that describe personal

[1] http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

events that occurred as opposed to entries with opinions and political commentary.

- **300 Yes/No questions and answers about the documents (6-12 per document)**
  - Human annotators will read the text and create the questions
  - They will also provide event bounding and identification annotated in the source documents (using only <EVENT> tag and eid attribute).
- NOTE: for testing the questions will not be given to participants until the organizers release the official results.

Crowd-sourced Test, as a novelty, in addition to the regular test, participants can opt to participate also in this social test. This is a testset created by the participants in an equitative way. To participate in this evaluation a participant is required to:

- Select or create an ascii plain document in English (any kind) that does not have a human TimeML annotation. Min. size: 120 words.
  - The recommended domains are: Wiki history and biographical (en.wikipedia.org), News (Wikinews, NTY, WSJ), narrative blogs (e.g., describing personal events) the later is very interesting to see if systems have basic capabilities in simple texts. Example: "I woke up this morning and went to school. There was my friend John sitting in the first row. Before taking a seat I hanged up my jacket. Then, I told John that this morning before coming to school I had a milkshake for breakfast and it was delicious"
- Create a set of ten temporal questions in QA tempeval format, including the document annotated with the events included in the questions.
- Get both the document and the question set approved by the organizers before the testing period.

Shared test in collaboration with Clinical and Timeline Tasks: Under construction

Tools and evaluation schedule

The participants are provided with a temporal QA system [UzZaman et al. 2012] that given a set of questions and a TimeML annotated document is capable of creating a timegraph [Miller&Schubert 1990] representing the temporal knowledge annotated, which is finally used to perform temporal reasoning and provide answers for the questions and give a score. TempEval-3 scoring tools can also be used to tune the systems.

Training/Dev period

Participants are given the questions and answers, and the related documents in TimeML. They will have access to the official training set and also the "social-training" set that they can optionally create in a collaborative effort.

Knowing the answers and the TimeML event annotations participants can fine tune their systems to perform temporal information understanding (TimeML annotation). In addition to being able to evaluate their TimeML performance over training data, using the provided QA system, with the questions and the output of their systems they can evaluate the temporal information understanding level of their approaches.

Testing period

The plain documents will be released and participants will be asked to annotate them with their systems.

- **Regular Test.** Only the plain text documents are provided (TE3-input format). The questions for each document and their answers will only be available after the evaluation. There will be no manual TimeML annotation for those documents.

- **Crowd-sourced Test.** During the regular testing period participant systems will be required to provide annotations for this crowd-sourced test that will be held at the same time as the regular one. Then the organizers will perform the evaluation with the question-set resulting of adding the questions from each participant.

The results as well as the question-sets will be released for participants to analyze their results and improve systems for future editions.

There will be a joint and separated analysis of results by official vs crowd-sourced test and further specific analysis like news vs wiki documents or intra vs inter-sentence questions.

Discussion

The evaluation results measure how far we are on temporal information understanding and event ordering applied to an end-user task rather than the exact TimeML annotation performance. Although the results report will not be as detailed and TimeML specific as in previous TempEvals, they will point out if the important temporal information in the text is captured by the current systems. That is to say if automated TimeML annotations from current systems are good enough to perform an extrinsic task (temporal QA).

Test questions cover aspects of the documents relevant for human readers, so indirectly we are measuring the temporal awareness of the important events.

Temporal QA is an end-user application that requires temporal information understanding. QA TempEval goes further than TREC factual questions about time (mainly 'when' questions). In this evaluation, systems are asked about questions that require temporal reasoning and understanding of temporal between events that are distant in the source document.

With a YES/NO question set it could be thought that a 50% F-measure is something average. However, if an answer cannot be deduced from the TimeML annotation, it is predicted as UNKNOWN. Therefore, if a system, does not annotate any temporal entity in the given plain text document all the answers will be UNKNOWN. Furthermore, the participants do not provide the answers directly, they provide the TimeML annotation. Since they do not know the questions in advance they cannot create a system that always returns 'yes' or 'no'. A correct answer thus reflects that system annotations are correct and complete enough to cover the temporal knowledge required to answer the question.

Some benefits that QA TempEval brings to the field include:

- **Annotation time is considerably reduced**. Although the participants task is the same (equivalent to task ABC) the evaluation does not require manual TimeML annotation by the organizers.
- **Reliability without expertise**: for a human, it is easier to fail an annotation (unless you have at least 3 annotators) than failing the answer for one temporal question. QA is also easier to review and less expertise is needed. Answering questions about events of a text is a task any person can do.
- **Analyze event relevance**. Documents are annotated without any bias or rule to select events. We assume that events picked up by annotators to ask questions about after reading the text should be naturally relevant. We will analyze if the events selected by annotators tend to be clearly bounded or ordered in time and if those could be considered as important by a separated set of annotators.
- **Analyze event distance factor in both the annotations and the results**. E.g., is there a trend for human annotators to select closer or distant events for questions? do systems answer better questions about closer events?
- **Easy to evaluate new domains**. Although participants can blame that there is no training for those, they all are in the same conditions so the evaluation is still fair. Results can be analyzed by domain so the influence of having more domain specific TimeML data.

It is worth recalling that TimeML was created with the purpose of allowing Temporal QA.

In order to compare participating systems to previous state-of-the-art systems, organizers will evaluate and provide scores for TIPSem and CAEVO systems.

References

J. F. Allen, “Maintaining knowledge about temporal intervals,” Communication ACM, vol. 26, no. 11, pp. 832–843, 1983.

J. Pustejovsky, J. M. Castao, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev, “TimeML: Robust Specification of Event and Temporal Expressions in Text.” in New Directions in Question Answering, M. T. Maybury, Ed. AAAI Press, 2003, pp. 28–34.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G Katz, J. Pustejovsky, "SemEval-2007 Task 15: TempEval Temporal Relation Identification" in Proceedings of the 4th International Workshop on Semantic Evaluations, Ed. ACL, 2007 pp.75-80.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky, “Semeval-2010 task 13: Tempeval 2,” in Proceedings of International Workshop on Semantic Evaluations (SemEval 2010), 2010.

UzZaman et al “Semeval-2013 task 1: Tempeval 3,” in Proceedings of International Workshop on Semantic Evaluations (SemEval 2013), 2013.

UzZaman, Llorens, and Allen. 2012. Evaluating Temporal Information Understanding with Temporal Question Answering. IEEE ICSC