Question Answering Temporal Evaluation (QA TempEval)

SemEval 2015 Task 5

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen and James Pustejovsky

Presented by: Nasrin Mostafazadeh

June 2015

TempEval Challenge

- The TempEval challenge: Framework for evaluating systems that automatically annotate documents in TimeML format (Pustejovsky et al., 2003).
- Annotating in TimeML involves:
  - Extracting temporal expressions (timexes)
  - Extracting events
  - Identifying semantic relations (including temporal relation known as TLINK).
- Previous TempEvals:
  - TempEval (2007)
  - TempEval-2 (2010)
  - TempEval-3 (2013)

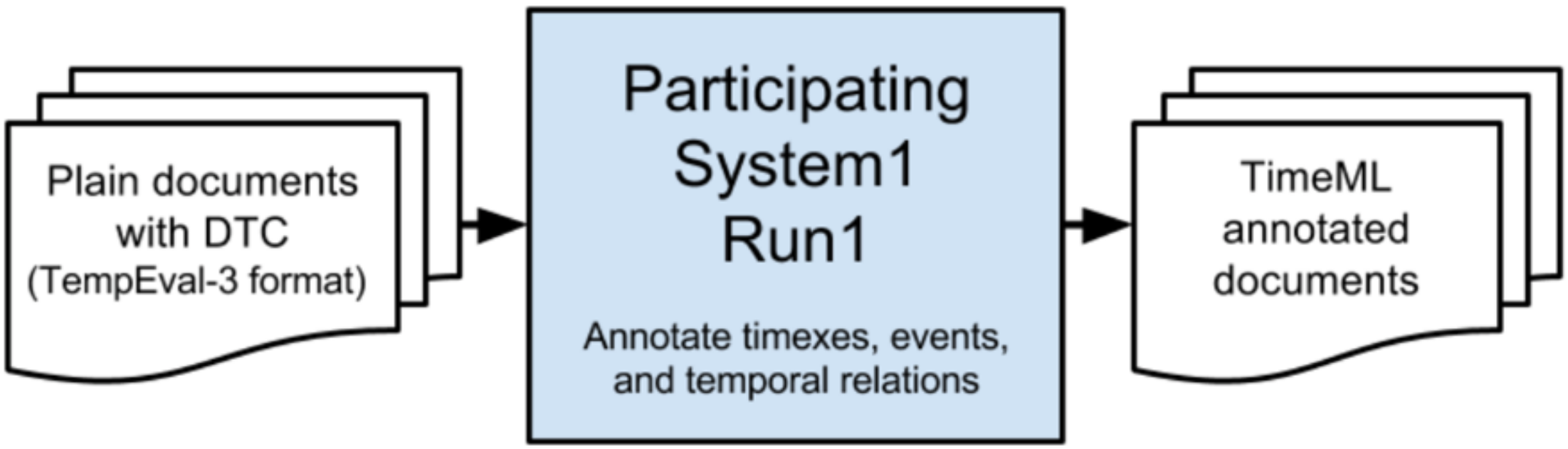
QA TempEval

- TimeML was originally developed to support research in complex temporal QA.
- Up to now: TempEval mainly focused on a more straightforward temporal information extraction (IE).
- **QA TempEval:**
  - Shifts the goal to QA (UzZaman et al., 2012), which:
    - Focuses on an end-user QA task.
    - Only important temporal information should be captured, not all.
    - Opposed to earlier corpus-based evaluation
- This evaluation is about the accuracy for answering the targeted questions:
  - Less expensive to develop the test-set (less expert time and effort)
  - More effective way to evaluate systems, as it involves understanding of the most important temporal information in a document

Overview of This Talk

- QA TempEval Task Description
- Test Data Creation
- QA Evaluation
- Participating Systems
- Results
- Conclusion

Task Description



The task for participant systems is equivalent to TempEval-3 task ABC.

Test Dataset Creation

- In QA TempEval, the dataset consists of question-sets and key documents.
- Dataset creation only requires the following:
  - Reading the document
  - Making temporal questions together with their correct answers
  - Bounding and identifying entities of the question in the text

Test Dataset (cont.)

- An example question and its corresponding annotated document:
  - **Question:**

    3|APW.tml|IS **ei21** AFTER **ei19**|Was he cited after becoming general?|yes

  - **APW.tml (KEY):**

    Farkas <event eid="**e19**">became</event> a general. He was <event eid="**e21**">cited</event>...

Test Dataset (cont.)

- The questions are **yes/no temporal questions** regarding any of the 13 Allen Interval relations holding between the two designated temporal entities.
- Annotators can ask any questions that comes naturally to a reader's mind
- However, the question creation is mainly focused **positive questions** (with *yes* answer).

Test Dataset (cont.)

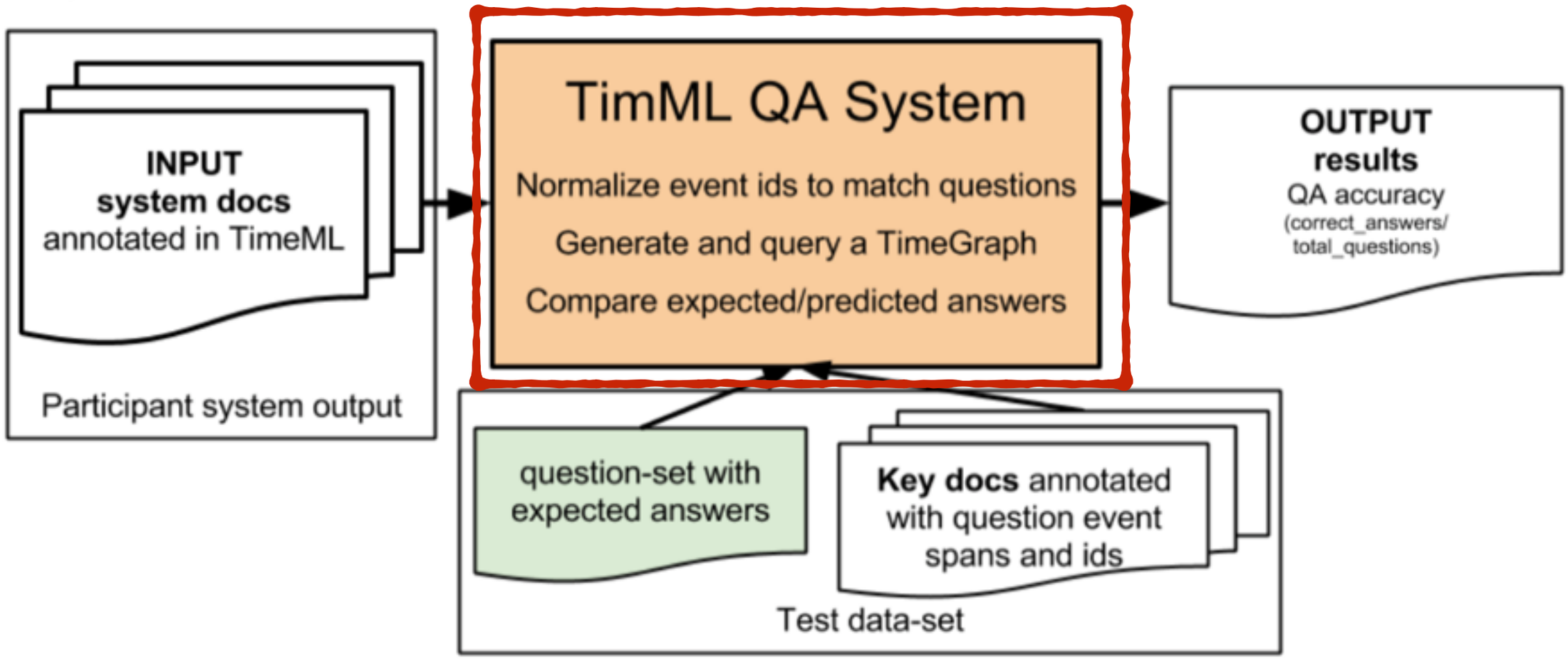
- QA TempEval is unique in expanding beyond the news genre:
  - News articles (Wikinews, WSJ, NYT)
  - Wikipedia articles (history, biographical)
  - Informal blog posts (narrative nature, describing personal stories):
- Human experts selected the documents and created the test dataset.
- The resulting question-set is then peer-reviewed by the human experts.

	docs	words	quest	yes	no	dist-	dist+
news	10	6920	99	93	6	40	59
wiki	10	14842	130	117	13	58	72
blogs	8	2053	65	65	0	30	35
total	28	23815	294	275	19	128	166

The statistics of the test-set

QA Evaluation

- Each system's annotations represent its temporal knowledge of the documents.
- The annotations of each system is then fed into a temporal QA system (UzZaman et al., 2012) that answers questions on behalf of the systems.



TimeML QA System

Given a system's TimeML annotated documents, the TimeML QA process consists of three main steps:

- **ID Normalization**: The system annotation IDs are aligned with the question IDs that are annotated in the key docs using the TempEval-3 normalization tool.
- **Timegraph Generation**: The normalized TimeML docs are used to build a graph of time points Here we use Timegraph (Gerevini et al., 1993) for computing *temporal closure* as proposed by Miller and Schubert (1990).
- **Question Processing**: Answering questions requires temporal information understanding and reasoning. Using Timegraph, the queries are converted to point-based queries.
  - For answering yes/no questions, we check the necessary point relations in Timegraph to verify an interval relation.

QA TempEval Participating Systems

- **Regular participants, optimized for task:**
  - **HITSZ-ICRC:** rule-based timex module, SVM (liblinear) for event and relation detection and classification
  - **hlt-fbk-ev1-trel1:** SVM, separated event detection and classification, without event coref
  - **hlt-fbk-ev1-trel2:** SVM, separated event detection and classification, with event coref
  - **hlt-fbk-ev2-trel1:** SVM, all predicates are events and classification decides, without event coref
  - **hlt-fbk-ev2-trel2:** SVM, all predicates are events and classification decides, with event coref
- **Off-the-Shelf Systems, not optimized on task:**
  - **CAEVO** (Chambers et al., 2014): Cascading classifiers that add temporal links with transitive expansion.
  - **ClearTK** (Bethard, 2013): A pipeline of machine-learning classification models, each of which have simple morphosyntactic annotation pipeline as feature set.
  - **TIPSemB** (Llorens et al., 2010): CRF-SVM model with morphosyntactic features
  - **TIPSem** (Llorens et al., 2010): TIPSemB + lexical (WordNet) and combinational (PropBank roles) semantic features

Results

- **Precision (P)** $= \frac{num\_correct}{num\_answered}$
- **Recall (R)** $= \frac{num\_correct}{num\_questions}$
- **F-measure (F1)** $= \frac{2*P*R}{P+R}$

Results over all domains:

	Measures			Questions	
System	P	R	F1	awd%	corr
HITSZ-ICRC	.54	.06	.12	.12	19
hlt-fbk-ev1-trel1	**.57**	.17	.26	.30	50
hlt-fbk-ev1-trel2	.47	.23	.31	.50	69
hlt-fbk-ev2-trel1	.55	.17	.26	.32	51
hlt-fbk-ev2-trel2	.49	**.30**	**.37**	**.62**	**89**
ClearTK	.59	.06	.11	.10	17
CAEVO	.56	**.17**	**.26**	**.31**	**51**
TIPSemB	.47	.13	.20	.28	38
TIPSem	**.60**	.15	.24	.26	45

Timex-Timex Reference Link (TREFL)

- **What will happen if we add a general time expression reasoner on top of systems?**
- The idea is to have a TREFL backbone to add all the implicit temporal relations between timexes:
  - TREFL resolves all time expressions and (when the relation is unambiguous) adds temporal relations between the time expressions(e.g., reason that 1999-01-12 is before 2015-06-06)
  - Any relations predicted by a classifier can be discarded if they are in-consistent with this TREFL backbone.
  - **Example**: assume that a system's TimeML annotation contains three timexes t1(1999), t2(1998-01-15), and t3(1999-08). Then TREFL adds 't2 BEFORE t1' and 't3 IS-INCLUDED t1'.

Timex-Timex Reference Link (TREFL)

	Measures			Questions	
System	P	R	F1	awd%	corr
HITSZ-ICRC	.58	.09	.15	.15	25
hlt-fbk-ev1-trel1	**.62**	.28	.38	.45	81
hlt-fbk-ev1-trel2	.55	.31	**.40**	.57	92
hlt-fbk-ev2-trel1	.61	.29	.39	.48	86
hlt-fbk-ev2-trel2	.51	**.34**	**.40**	**.67**	**99**
ClearTK (TREFL not applied because of its TLINK format)					
CAEVO	.60	.21	.32	.36	63
TIPSemB	.64	.24	.35	.37	70
TIPSem	**.68**	**.27**	**.38**	**.40**	**79**

Results with TREFL

	Measures			Questions	
System	P	R	F1	awd%	corr
HITSZ-ICRC	.54	.06	.12	.12	19
hlt-fbk-ev1-trel1	**.57**	.17	.26	.30	50
hlt-fbk-ev1-trel2	.47	.23	.31	.50	69
hlt-fbk-ev2-trel1	.55	.17	.26	.32	51
hlt-fbk-ev2-trel2	.49	**.30**	**.37**	**.62**	**89**
ClearTK	.59	.06	.11	.10	17
CAEVO	.56	**.17**	**.26**	**.31**	**51**
TIPSemB	.47	.13	.20	.28	38
TIPSem	**.60**	.15	.24	.26	45

Results without TREFL

Conclusion

- Systems are still far from **deeply understanding temporal aspects of NL** (QA Recall ~30%)!

Datasets and Results Available Online!

Task webpage, containing Data, Tools, and Results:

http://alt.qcri.org/semeval2015/task5/