Overview of the Fourth Recognising Textual Entailment Challenge

Danilo Giampiccolo (coordinator, CELCT)
Hoa Trang Dan (NIST)

Ido Dagan (Bar Ilan University)
Bill Dolan (Microsoft Research)
Bernardo Magnini (FBK-irst)

Textual Entailment

Textual entailment is a directional relation between two text fragments –the entailing text, *called t(ext)*, and the entailed text, called *h(ypothesis),* so that a human being, with common understanding of language and common background knowledge, can infer that ***h* is most likely true on the basis of the content of *t*.**

What was new in RTE 4

- RTE was **organised jointly by NIST** and CELCT and proposed as a **track of the Text Analysis Conference**.

- **Three-way annotation**: introduced by NIST as a pilot task at the Workshop for Paraphrasing and Textual Entailment, ACL 2007, was proposed in the main task, where the systems were required to make a further *distinction* between pairs where *the entailment does not hold* because the content of H *is contradicted* by the content of T, and pairs where **the entailment cannot be determined** because the truth of H cannot be verified on the basis of the content of T.

Definition of the task

Given two text snippets - t and h - the system must decide whether:

- 3 way task:
  - **T entails H** - in which case the pair is marked as ENTAILMENT
  - **T contradicted H** in which case the pair is marked as CONTRADICTION
  - **The truth of H could not be determined on the basis of T,** in which case the pair is marked as **UNKNOWN**
- 2-way task:
  - **T entailed H,** in which case the pair is marked as **ENTAILMENT**
  - T does not entailed H in which case the pair is marked as **NO ENTAILMENT**

Examples

- **YES** (entailment holds):
  - T*: Spencer Dryden, the drummer of the legendary American rock band Jefferson Airplane, passed away on Tuesday, Jan. 11. He was 66. Dryden suffered from stomach cancer and heart disease.*
  - H: *Spencer Dryden died at 66.*
- **CONTRADICTION** (T contradicts H):
  - T: *Lower food prices pushed the UK's inflation rate down to 1.1% in August, the lowest level since 1963. The headline rate of inflation fell to 1.1% in August, pushed down by falling food prices.*
  - H: *Food prices are on the increase.*
- **UNKNOWN** (not possible to determine the entailment)
  - T: *Four people were killed and at least 20 injured when a tornado tore through an Iowa boy scout camp on Wednesday, where dozens of scouts were gathered for a summer retreat, state officials said.*
  - H: *Four boy scouts were killed by a tornado.*

The Data Set

- No development set this year
- 1000 *t-h* pairs (IE and IR proved to be more difficult)
  - 300 IE
  - 300 IR
  - 200 QA
  - 200 SUM
- Longer *t,* with respect to RTE3
- Distribution according the entailment:
  - 50% ENTAILMENT
  - 35% UNKNOWN
  - 15% CONTRADICTION

Text sources

The same as last year:

- Output data (both correct and incorrect) of Web-based systems
- Input data publicly released by official competitions
- Freely available sources such as WikiNews and Wikipedia

Pair collection: IE setting

- Inspired by Information Extraction, where texts and structured templates are turned into *t-h* pairs.

- Simulates the need of IE systems to recognize that the given text entails the semantic relation that is expected to hold between the candidate template slot fillers.

Pair collection: QA setting

- From Question-Answer pairs to *t-h* pairs:
  - An answer term of the expected answer type is picked from the answer passage.
  - The question is turned into an affirmative sentence plugging in the answer term.
  - *t-h* pairs are generated, using the affirmative sentences as hypotheses and the original answer passages as texts-

- This process simulates the need of a QA system to verify that the retrieved passage text entails the provided answer.

Pair collection: SUM setting

- Given sentence pairs from the output of multi-document summarization systems, hypotheses are generated by removing sentence parts:
  - for positive examples, the hypothesis is simplified by removing sentence parts, until it is fully entailed by T. Negative examples – i.e. where the entailment does not hold- are produced in a similar way, i.e. taking away parts of T so that the final information contained in H either contradicts the content of T, or is not enough to determine the entailment.
- This process simulates the need of a summarization system to identify information redundancy, which should be avoided in the summary.

Evaluation measures

- Automatic evaluation:
  - **_Accuracy_** (main evaluation measure): percentage of correct judgments against the Gold Standard
  - ***Average precision*** *(for systems which returned a confidence score)*: average of the system's precision values at all points in the ranked list in which recall increases, that is at all points in the ranked list for which the gold standard annotation is YES. In the case of three-way judgment submissions the pairs tagged as CONTRADICTION and UNKNOWN were conflated and retagged as NO ENTAILMENT.

Participants

- **Participants at RTE4: 26**
  - RTE1 18
  - RTE2 23
  - RTE3 26
- **Provenance**
  - USA: 9
  - EU: 13
  - ASIA: 4
- **Participants for tasks**
  - 8 at 3-way only
  - 13 at 2-way only
  - 5 at both

Results: Average Accuracy

THREE-WAY TASK									
3-way					2-way				
Overall	**IE**	**SUM**	**IR**	**QA**	**Overall**	**IE**	**SUM**	**IR**	**QA**
0.51	0.4536	0.5080	0.5381	0.4527	0.5641	0.5211	0.5873	0.6178	0.5250

TWO-WAY TASK				
Overall	**IE**	**SUM**	**IR**	**QA**
0.57 **(0.61 at RTE3)**	0.521844 (0.52)	0.602556 (0.58)	0.615182 (0.66)	0.525778 (0.71)

Results: BEST RESULTS

RANKING					
THREE-WAY TASK				TWO-WAY TASK	
3-W		2-W			
UAIC20081	**0.685**	UAIC20081	0.72	Icc1	**0.746**
OAQA1	0.616	OAQA1	0.688	UAIC20081	0.721
DFKI1	0.614	DFKI1	0.687	DFKI3	0.706
DFKI2	0.606	DFKI2	0.67	DFKI2	0.699
QUANTA1	0.588	QUANTA1	0.664	DFKI1	0.672
DFKI3	0.56	DFKI3	0.633	QUANTA1	0.659
UMD1	0.556	UMD1	0.619	QUANTA2	0.623
UMD2	0.556	UMD2	0.617	DLSIUAES1	0.608

Resources

- WordNet, Extended WordNet, Extended WordNet Knowledge Base
- DIRT
- FrameNet, ProBank, VerbNet
- Entailment pairs
- Corpora (e.g. for estimating IDF)
- Antonym expressions
- Gazzetteers
- Wikipedia

Methods

- **Lexical similarity**
  - Word overlap, Edit distance, etc.
- **Alignment based on syntactic representations**
  - Tree Edit Distance, tree kernels
- **Alignment based on committees**
- **Transformation based approaches**
  - Probabilistic setting
- **Individuate contradictions**
- **Machine learning**
  - Classifiers take the final decision
- **Logical inferences**
  - Ontology based reasoning
- **Combining specialized entailment engines**
  - Voting, etc.

Conclusion

- RTE-4 organization moved to NIST with CELCT involved as coordinator
- High level of maturity and diffusion of textual entailment
- 3-way evaluation has been introduced