RTE-5 MAIN TASK GUIDELINES

(Also see general TAC 2009 policies and guidelines at http://www.nist.gov/tac/2009/)

The Recognizing Textual Entailment (RTE) challenge is an annual exercise that provides a framework for evaluation of entailment systems, and promotes international research in this area. In this evaluation exercise systems are given T-H pairs and must determine whether T entails H.

We define textual entailment as a directional relationship between a pair of text fragments, which we call the "Text" (T) and the "Hypothesis: (H). We say that:

T entails H, denoted by T→H, if a human reading T would infer that H is most likely true.

This definition of entailment is based on (and assumes) common human understanding of language as well as common background knowledge.

For example, given assumed common background knowledge of the business news domain and the following text:

> *T1 Internet media company Yahoo Inc. announced Monday it is buying Overture Services Inc. in a $1.63-billion (U.S.) cash-and-stock deal that will bolster its on-line search capabilities.*,

the following hypotheses are entailed:

- *H1.1 Yahoo bought Overture*
- *H1.2 Overture was acquired by Yahoo*
- *H1.3Overture was bought*
- *H1.4 Yahoo is an internet company*

If H is not entailed by T, there are two possibilities:

1) H contradicts T
2) the information in H cannot be judged as TRUE on the basis of the information contained in T.

For example, the following hypotheses are contradicted by T1 above:

- *H1.5 Overture bought Yahoo*
- *H1.6 Yahoo sold Overture*

While the following ones cannot be judged on the basis of T1 above:

- *H1.7 Yahoo manufactures cars*
- *H1.8 Overture shareholders will receive $4.75 cash and 0.6108 Yahoo stock for each of their shares.*

TASK DESCRIPTION

Textual entailment recognition is the task of deciding, given a T-H pair, whether T entails H.

The main task consists of two sub-tasks:

1. The *three-way* RTE task, where the system must decide whether:
   - T *entails H* - in which case the pair will be marked as ENTAILMENT
   - T c*ontradicts* H - in which case the pair will be marked as CONTRADICTION
   - The *truth of H cannot be determined on the basis of T* - in which case the pair will be marked as UNKNOWN
2. The *two-way* RTE task is to decide whether:
   - T *entails H* - in which case the pair will be marked as ENTAILMENT
   - T *does not entail* H - in which case the pair will be marked as NO ENTAILMENT

When entailment holds: general rules

a) The hypothesis must be fully entailed by the text. The judgment cannot be ENTAILMENT if the hypothesis includes parts that cannot be inferred from the text.

b) If no major context is provided, the entities and the events mentioned in H and T are always supposed to be the same. So, considering the following example:

T2: *Yesterday 30 people were killed in a train accident near London.*
H2: *27 people died in a train accident.*

we must suppose that both H and T refer to the same train accident, giving a contradictory report about the number of casualty. More details are discussed below.

c) Prior knowledge or presupposition of common knowledge is always supposed, for instance: a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. However, entailment must be judged considering the content of T **AND** COMMON KNOWLEDGE **TOGETHER**, and NEVER ON THE BASIS OF COMMON KNOWLEDGE ALONE. For example:

T3: The recovery of the capsule, which carried astronaut Virgil "Gus" Grissom on a brief suborbital flight on July 21, 1961, took place on the 30th anniversary of mankind's first moon landing.
H3: The moon was first touched by mankind in 1969.

is UNKNOWN, even if probably common readers, referring to their background knowledge, know that the information in H is correct (but non inferable from T).

d) Entailment is assumed even if it is just very probable rather than certain. For example, *John purchased the book* should entail *John paid for the book*; even if it might theoretically be possible to buy something without paying for it. By the same token, *Mary criticized the proposal* should NOT entail *Mary rejected the proposal* unless there is a strong reason to believe that indeed Mary rejected the idea.

Further examples:

T4: Russia doesn't excel in auto manufacturing, but it is tops in other areas. The country's great wealth in oil, natural gas, and metals, in addition to generating revenue for the government, is creating a new middle class that craves European, Japanese and American cars.
H4: European cars sell in Russia.
Assessment: ENTAILMENT

e) Entailment is a directional relation. The hypothesis must be fully entailed from the given text, but the text need not be entailed from the hypothesis. In that perspective, it might help to first read the hypothesis and understand what it states, and only then read the text and see if it has sufficient information to entail the hypothesis.

f) People mentioned in T and H should be treated as co-referent in the absence of clear countervailing evidence. For example, below it should be assumed that the two expressions "a woman" refer to the same woman.

T5: Passions surrounding Germany's final match at the Euro 2004 soccer championships turned violent when a woman stabbed her partner in the head because she didn't want to watch the game on television.
H5: A woman passionately wanted to watch the soccer championship.
Assessment: CONTRADICTION

Whether to regard a text and hypothesis as describing the same event is more subtle. If two descriptions appear overlapping, rather than completely unrelated, by default assume that the two passages describe the same context, and contradiction is evaluated on this basis. For example, if there are details that seem to make it clear that the same event is being described, but one passage says it happened in 1985 and the other 1987, or one passage says two people met in Greece, and the other in Italy, then you should regard the two as a contradiction. Below, it seems reasonable to regard "a ferry collision" and "a ferry sinking" as the same event, and then the reports make contradictory claims on casualties:

T6: Rescuers searched rough seas off the capital yesterday for survivors of a ferry collision that claimed 28 lives, as officials blamed crew incompetence for the accident.
H6: 100 or more people lost their lives in a ferry sinking.
CONTRADICTION

In other circumstances, it may be most reasonable to regard the two passages as describing different events, if the descriptions are very different.

Contemporaneousness of facts and actions

If T and H refer to the same event, and there is no countervailing evidence, the distinction between past and present verb tenses MUST BE IGNORED, and the actions and facts presented in T/H must be considered as CONTEMPORANEOUS, i.e. happening at the same time.

For example:

T7: *Yahoo acquired Overture.*
H7: *Yahoo is buying Overture.*

is assessed as ENTAILMENT, as, regardless the different tenses, there is no contradiction, and the action of buying described in T7 must be considered contemporaneous to the H7.

By the same token,

T8: Cuban Leader Fidel Castro sent a letter to United Nations Secretary-General Kofi Annan assuring him that Cuba will follow anti-terrorism treaties in the wake of last month's terrorist attacks against New York and Washington.
H8: Castro visits the UN.

is assessed as CONTRADICTION because, assuming that:

- the entities and the event mentioned in T and H co-refer
- tense difference must be ignored
- the actions (sending a letter and visiting the UN) must be considered contemporaneous,

it is highly unlikely that Castro visits and sends a letter to the UN at exactly the same time.

When entailment does not hold: CONTRADICTION vs UNKNOWN

In order to help distinguish between a CONTRADICTION and an UNKNOWN judgment, the following hints must be considered:

a) H contradicts T if assertions in the hypothesis appear to directly refute or show portions of the text to be false/wrong, if the hypothesis were taken as reliable.

T20: Jennifer Hawkins is the 21-year-old beauty queen from Australia.
H20: Jennifer Hawkins is Australia's 20-year-old beauty queen.
Assessment: CONTRADICTION

T21: In that aircraft accident, four people were killed: the pilot, who was wearing civilian clothes, and three other people who were wearing military uniforms.
H21: Four people were assassinated by the pilot.
Assessment: CONTRADICTION

b) A text and hypothesis reporting contradictory statements is marked as a CONTRADICTION, if the reports are stated as facts (these cases can be seen as embedded contradictions)

T22: That police statement reinforced published reports, that eyewitnesses said Menezes had jumped over the turnstile at Stockwell subway station and was wearing a padded jacket, despite warm weather.
H22: ITV News reported that Menezes walked casually into the subway, wearing a light denim jacket.
Assessment: CONTRADICTION

DATASET FORMAT

The dataset of Text-Hypothesis pairs is collected by human annotators and consists of three subsets which correspond to different application settings: Information Extraction (IE), Information Retrieval (IR), and Question Answering (QA).

The dataset is formatted as an XML file, as follows:

<pair id="id_num" entailment="ENTAILMENT|
CONTRADICTION|UNKNOWN" task="IE|IR|QA">

<t>the text...</t>
<h>the hypothesis...</h>
</pair>

Where:

- each T-H pair appears within a single <pair> element.
- the element <pair> has the following attributes:
  - id, a unique numeral identifier of the T-H pair.
  - task, the acronym of the application setting from which the pair has been generated: "IR","IE", or "QA".
  - entailment (in the gold standard only), the gold standard entailment annotation, being either "ENTAILMENT", "CONTRADICTION" or "UNKNOWN"
- the element <t> (text) has no attributes, and it may be made up of one or more sentences.
- the element <h> (hypothesis) has no attributes, and it usually contains one simple sentence.

IMPORTANT: Participants must submit the "Agreement Concerning Dissemination of TAC Results" before they can use their TAC 2009 team ID and password to download any data. A link to the agreement form and instructions for submitting forms are found at the TAC 2009 tracks home page:
http://www.nist.gov/tac/2009/

RESULT SUBMISSION

Teams can participate in either 2- or 3-way task, or both. No partial submissions are allowed, i.e. the submission must cover the whole dataset. Each team is allowed to submit up to 6 runs (up to 3 runs for each task). This allows teams who attempt both 3-way and 2-way classification to optimize/train separately for each task. Teams that participate in the 3-way task and do not have a separate strategy for the 2-way task (other than to automatically conflate CONTRADICTION and UNKNOWN to NO ENTAILMENT), should not submit separate runs for the 2-way task, because runs for the 3-way task will automatically be scored for both the 3-way task and the 2-way task.

Each run may optionally rank all the T-H pairs in the test set according to their entailment confidence (in decreasing order from the most certain entailment to the least certain). The more the system is confident that T entails H, the higher the ranking is. A perfect ranking would place all the pairs for which T entails H, before all the pairs for which T does not entail H. Because the evaluation measure for confidence ranking applies only to the 2-way classification task, in the case of three-way runs the pairs tagged as CONTRADICTION and UNKNOWN will be conflated and automatically re-tagged as NO ENTAILMENT for scoring purposes.

Runs will be submitted using a password-protected online submission form on the RTE web page (http://www.nist.gov/tac/2009/RTE/). The link to the submission form will be posted when the test dataset is released.

At the time of submission, each team will be asked to fill out the form stating:

- Whether the submission is for the three-way task or the two-way task
- Whether the pairs are ranked in order of entailment confidence
- A number (1-3) for the run, used to differentiate between the team's runs for the task

NB: Analyses of the test set (either manual or automatic) should not impact in any way the design and tuning of systems that publish their results on the RTE-4 test set. We regard it as acceptable to run automatic knowledge acquisition methods (such as synonym collection) specifically for the lexical and syntactic constructs that will be present in the test set, as long as the methodology and procedures are general and not tuned specifically for the test data. In any case, participants are asked to report about any process that was performed specifically for the test set.

RESULT SUBMISSION FORMAT

Results will be submitted as one file per run. Each submitted file must be a plain ASCII file with one line for each T-H pair in the test set, in the following format:

- pair_id judgment

Where:

- pair_id is the unique identifier of each T-H pair, as it appears in the test set
- judgment for the three-way task is one of: ENTAILMENT, CONTRADICTION, or UNKNOWN
- judgment for the two-way task is one of: ENTAILMENT or NO ENTAILMENT

If the run includes confidence ranking, then the pairs in the file should be ordered by decreasing entailment confidence: the first pair should be the one for which the entailment is most certain, and the last pair should be the one for which the entailment is least likely. Thus, in a ranked run, all the pairs classified as ENTAILMENT are expected to appear before all the pairs that are classified as NO ENTAILMENT (for the two-way task) or CONTRADICTION or UNKNOWN (for the three-way task).

EVALUATION MEASURES

The evaluation of all submitted runs will be automatic. The judgments (classifications) returned by the system will be compared to those manually assigned by the human annotators (the Gold Standard). For the two-way task, a judgment of "NO ENTAILMENT" in a submitted run is considered to match either "CONTRADICTION" or "UNKNOWN" in the Gold Standard. The percentage of matching judgments will provide the accuracy of the run, i.e. the fraction of correct responses.

As a second measure, an Average Precision score will be computed for systems that provide as output a confidence-ranked list of all test examples. This measure evaluates the ability of systems to rank all the T-H pairs in the test set according to their entailment confidence (in decreasing order from the most certain entailment to the least certain). The more the system is confident that T entails H, the higher the ranking is. A perfect ranking would place all the pairs for which T entails H, before all the pairs for which T does not entail H. Average precision is a common evaluation measure for system rankings, and is computed as the 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 ENTAILMENT. More formally, it can be written as follows:

1/R * sum for i=1 to n (E(i) * #-entailing-up-to-pair-i/i)

where n is the number of the pairs in the test set, R is the total number of ENTAILMENT pairs in the Gold Standard, E(i) is 1 if the i-th pair is marked as ENTAILMENT in the Gold Standard and 0 otherwise, and i ranges over the pairs, ordered by their ranking. As average precision is relevant only for a binary annotation, in the case of three-way judgment submissions the pairs tagged as CONTRADICTION and UNKNOWN will be conflated and re-tagged as NO ENTAILMENT.

ABLATION TESTS

An ablation test consists of removing one module at a time from a system, and rerunning the system on the test set with the other modules, except the one tested. Comparing the results to those obtained by the system as a whole, it is possible to assess the practical contribution given by each single module.

In order to better understand the relevance of the knowledge resources used by RTE systems, and evaluate the contribution of each of them to the systems' performances, ablation tests for major knowledge resources will be required for those systems that employ these resources, to be submitted together with the system runs.

Should the knowledge resources used appear to be too numerous to run ablation tests on all of them, it will be asked to test at least three of them, specifically those which are thought to have the greatest impact on the overall performance.

IMPORTANT DATES

Main Development Set Release:	29 May 2009
Main Test Set Release:	2 September 2009
Submissions:	9 September 2009
Release of individual evaluated results:	18 September 2009