7th TEXTUAL ENTAILMENT CHALLENGE @ TAC 2011

MAIN TASK and NOVELTY DETECTION SUBTASK
Task Guidelines

1. INTRODUCTION

The Recognizing Textual Entailment (RTE) task consists of developing a system that, given two text fragments, can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other text. Since it inception in 2005, RTE has enjoyed a constantly growing popularity in the NLP community, as it seems to work as a common framework in which to analyze, compare and evaluate different techniques used in NLP applications to deal with semantic inference, a common issue shared by many NLP applications.

After the first three highly successful PASCAL RTE Challenges campaigns held in Europe, RTE became a track at the Text Analysis Conference (TAC 2008), bringing it together with communities working on NLP applications. The interaction has provided the opportunity to apply RTE systems to specific application settings and move them towards more realistic scenarios. In particular, the RTE-5 Pilot Search task represented a step forward, as for the first time textual entailment recognition was performed on a real text corpus. Furthermore, it was set up in the Summarization setting, attempting to analyze the potential impact of textual entailment on a real NLP application.

In RTE-6, the traditional Main task was replaced by the task of Recognizing Textual Entailment within a corpus. The new RTE-6 Main task, situated in the Summarization application setting, was a close variant of the Pilot Search task in RTE-5 and had two goals, namely i) to advance the state of the art in RTE, by proposing a data set which reflects the natural distribution of entailment in a corpus and presents all the problems that can arise while detecting textual entailment in a natural setting and ii) to further explore the contribution that RTE engines can make to Summarization applications. In fact, in a general summarization setting, correctly extracting all the sentences entailing a given candidate statement for the summary (similar to Hypotheses in RTE) corresponds to identifying all its mentions in the text, which is useful to assess the importance of that candidate statement for the summary and, at the same time, to detect those sentences which contain redundant information and should probably not be included in the summary. Furthermore, if automatic summarization is performed in the Update scenario (where systems are required to write a short summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles) it is important to distinguish between novel and non-novel information. In such a setting, RTE engines which are able to detect the novelty of H's can help Summarization systems filter out non-novel sentences from their summaries.

In order to ensure the continuity with the previous campaign and allow the participants to get acquainted with the novelties introduced for the first time in the two RTE-6 tasks situated in the Summarization scenario[1], the same tasks are proposed again in RTE-7 without significant changes, specifically:

- **Main Task - *Recognizing Textual Entailment within a Corpus*:** Given a corpus, a hypothesis H, and a set of "candidate" sentences retrieved by Lucene from that corpus for H, RTE systems are required to identify all the sentences that entail H among the candidate sentences.

[1] See (Bentivogli et al., 2010).

- ***Novelty Detection* subtask:** Based on the Main task, the subtask is focused on Novelty Detection, which means that RTE systems are required to judge whether the information contained in each H is novel with respect to (i.e., not entailed by) the information contained in the corpus. If entailing sentences are found for a given H, it means that the content of the H is not new; in contrast, if no entailing sentences are detected, it means that information contained in the H is novel. Although the Novelty Detection task has the same structure as the Main task, it is separated out as a subtask to allow participants to optimize their RTE engines differently (i.e., for novelty detection). Systems' outputs will have the same format as for the Main task – i.e. no additional type of information is needed – but will be scored differently, to better reflect the goal of novelty detection.

This document provides a definition of both the Main task and Novelty Detection subtask, together with a description of the data set and the instructions on how to take part in the exercise.

2. RECOGNIZING TEXTUAL ENTAILMENT WITHIN A CORPUS

According to the standard definition [Dagan et al., 2006], Textual Entailment is defined as a directional relationship between two text fragments, termed Text (T) and Hypothesis (H). It is said that:

T ENTAILS H IF, TYPICALLY, 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 background knowledge; in fact, for textual entailment to hold it is required that:

TEXT AND KNOWLEDGE ENTAIL H, **BUT** KNOWLEDGE ALONE CANNOT ENTAIL H

This means that H may be entailed by incorporating some prior knowledge that would enable its inference from T, but it should not be entailed by that knowledge alone. In other words, it is not allowed to validate H's truth regardless of T.

The traditional RTE Main task, which was carried out in the first five RTE challenges, consisted of making entailment judgments over isolated T-H pairs. In such a framework, both Text and Hypothesis were artificially created in a way that they did not contain any references to information outside the T-H pair. As a consequence, the context necessary to judge the entailment relation was given by T, and only language and world knowledge were needed, while reference knowledge was typically not required.

In contrast, the task of Recognizing Textual Entailment within a corpus, which has been proposed as the Main task since RTE-6[2], consists of finding all the sentences in a set of documents that entail a given Hypothesis. In such a scenario, both T and H are to be interpreted in the context of the corpus, as they rely on explicit and implicit references to entities, events, dates, places, situations, etc. pertaining to the topic[3]. Moreover, such a task requires the retrieval of entailing sentences only, and the entailment judgment may be seen as a two-way decision between "yes" and "no" entailment.

3. RTE-7 MAIN TASK

3.1 TASK DESCRIPTION

The RTE-7 Main task is situated in the Summarization application setting and replicates the RTE-6 Main task, with some minor changes.

[2] This new kind of Textual Entailment task, called Search task, was first introduced as a pilot task in RTE-5 (see Bentivogli et al., 2009b).

[3] For an analysis of the relevance of discourse phenomena in Textual Entailment see (Bentivogli et al., 2009a).

In the RTE-7 Main task, given a corpus, a hypothesis H, and a set of "candidate" entailing sentences for that H retrieved by Lucene from the corpus, RTE systems are required to identify all the sentences that entail H among the candidate sentences.

Similarly to what happened in RTE-6 Main task, it must be noted that:

- a preliminary Information Retrieval filtering phase is performed by the organizers using Lucene, in order to select for each H a subset of candidate entailing sentences to be judged by the participating systems;
- some of the H’s have no entailing sentences.

The example below presents a hypothesis referring to a given topic and some of the entailing sentences found in the set of candidate sentences:

```
<H_sentence>Lance Armstrong is a Tour de France winner.</H_sentence>
<text doc_id="AFP_ENG_20050824.0557" s_id="1" evaluation="YES">Claims by a French newspaper that seven-
    time Tour de France winner Lance Armstrong had taken EPO were attacked as unsound and unethical by the
    director of the Canadian laboratory whose tests saw Olympic drug cheat Ben Johnson hit with a lifetime
    ban.</text>
<text doc_id="AFP_ENG_20050824.0557" s_id="2" evaluation="YES">L'Equipe on Tuesday carried a front page
    story headlined "Armstrong's Lie" suggesting the Texan had used the illegal blood booster EPO (erythropoeitin)
    during his first Tour win in 1999.</text>
<text doc_id="AFP_ENG_20050831.0529" s_id="1" evaluation="YES">The exploits of seven-times Tour de France
    champion Lance Armstrong, who is alleged to have used the banned blood booster EPO (erythropoietin) in 1999,
    are also down to the use of other banned substances according to one expert.</text>
<text doc_id="AFP_ENG_20050831.0529" s_id="3" evaluation="YES">Armstrong, who retired after his seventh
    yellow jersey victory last month, has always denied ever taking banned substances, and has been on a major
    defensive since a report by French newspaper L'Equipe last week showed details of doping test results from the
    Tour de France in 1999.</text>
<text doc_id="APW_ENG_20050823.0684" s_id="1" evaluation="YES">French sports daily L'Equipe reported
    Tuesday that Lance Armstrong used the performance-enhancing drug EPO to help win his first Tour de France in
    1999, a report the seven-time Tour winner vehemently denied.</text>
```

In order to clarify the process necessary to interpret the entailment relation in the RTE-7 Main task, Appendix A presents the guidelines which have been followed by the annotators when creating the data set. Note that the T’s are interpreted in their context, taking into account all the discourse references. For instance, the second sentence in the example above (doc_id="AFP_ENG_20050824.0557" s_id="2"） is considered an entailing sentence because from its context it can be seen that “*the Texan*” and *“Tour”* refer to “*Lance Armstrong*” and *“Tour de France”*, mentioned earlier in the discourse.

3.2 Data Set Description

The RTE-7 Main data set is based on the data created for the TAC 2008 and 2009 Update Summarization task, which consist of a number of topics, each containing two sets of documents, namely i) Cluster A, made up of the first 10 texts in chronological order (of publication date), and ii) Cluster B, made up of the last 10 texts.

The RTE-7 data set is composed of 20 topics, 10 used for the Development Set and 10 for the Test Set.

Note that RTE-7 participants must *not* process the original TAC Summarization data. Instead, the data must be regarded as blind, until the RTE-7 competition is complete.

For each topic, the RTE-7 Main task data consist of:

a) A number of Hypotheses (between 25 and 45) referring to the topic. H’s are standalone sentences taken from the TAC Update Summarization corpus – i.e. both Cluster A and

Cluster B documents[4]. When needed, minor syntactic and morpho-syntactic changes have been made with respect to the original sentences, from which the H’s are taken, to produce grammatically correct standalone sentences. Moreover, all the discourse references have been resolved.

b) A set of 10 documents, corresponding to the Cluster A corpus.

c) For each H, a list of up to 100 candidate entailing sentences from the Cluster A corpus and their location in the corpus. The candidate sentences are the 100 top-ranked sentences retrieved by Lucene, using H verbatim as the search query [5]. The Lucene ranking score will be provided to the participants as supplementary information regarding the preliminary IR phase.

Note that while only the subset of the candidate entailing sentences must be judged for entailment, these sentences are not to be considered as isolated texts. Rather, the entire Cluster A corpus, to which the candidate entailing sentences belong, is to be taken into consideration in order to resolve discourse references and appropriately judge the entailment relation (for more information see Appendix A). Also note that:

a) Contradictions are not considered in this task, and thus the entailment judgment choice must be between “yes” and “no” entailment.

b) A number of H’s have no entailing sentences.

c) Not all documents in the corpus contain entailing sentences.

d) Each document contains data which can be used in the entailment judgment, namely (i) the document ID - which encodes the date of publication, (ii) a headline, and sometimes (iii) an explicit dateline, for example:

```
<DOC doc_id="AFP_ENG_20050824.0557" type="story">
<HEADLINE>Top Canadian drug tester attacks Armstrong drug 'expose'</HEADLINE>
<DATELINE>MONTREAL, Aug 24</DATELINE>
```

e) Documents may contain non-informative sentences which, if found among the candidate entailing sentences, must be judged as NO entailment. Five examples are:

“_____” ; “On the net:” ; “1994” ; “New York:” ; “(Begin optional trim)”

f) T and H are naturally anchored to the publication date of the document from which they are taken (as for where to find the information about the publication date, see Item d. above and Item a. in 3.3). This must be taken into account while interpreting T and H verb tenses, since verb tenses are intrinsically deictic and depend on their anchor time (for more detail, see Bentivogli et al., 2009a).

In order to download the RTE-7 data (the Development Set and the Test Set) please submit the following user agreements to TAC 2011:

1. Agreement Concerning Dissemination of TAC Results
2. AQUAINT-2 Agreement

A link to all the user agreements can be found at the TAC 2011 home page (http://www.nist.gov/tac/2011/).

[4] In order to be as consistent as possible with the SUM scenario, some of the H’s are based on the content of Cluster B automatic summaries produced by the 3 best scoring systems participating in the TAC 2008 and 2009 Update Summarization task. An additional number of H’s have been created directly from Cluster A corpus sentences, to obtain a sufficient number of entailing sentences necessary for the RTE task.

[5] Results obtained on the whole RTE-5 Search dataset and on three topics of the RTE-6 Development Set show that, when the first 100 top-ranked sentences for each H are taken as candidates, Lucene achieves a recall of about 0,80. This implies that about 20% of entailing sentences, present in the corpus but not retrieved by Lucene, get lost.

3.3 Data Set and Submission Format

The whole data set is in XML format.

Development Set

The following items will be distributed as the Development Set:

1) The Development Set Gold Standard (see Example 1);

2) For each topic:

- <u>*Item A*</u>: a list of hypotheses (see Example 2). The element <ref> contains the corpus sentence from which the H is taken. The sentence itself is given for documentation purposes only and must not be used as additional data to deduce entailment. Its attribute “doc_id” indicates the ID of the document in which the sentence is found. This attribute encodes the publication date of the document and can be used to time anchor the hypothesis (e.g. *AFP_ENG_20050728.0294*, where *20050728* is the publication date in the format *YYYYMMDD*).
- <u>*Item B*</u>: for each hypothesis H, the list of the ID numbers of Cluster A candidate sentences to be judged for entailment (see Example 3). Note that for each candidate sentence its Lucene ranking score is given.
- <u>*Item C:*</u> the set of Cluster A documents for that topic.

<u>Example 1: Gold Standard</u>

```
<entailment_corpus>
   <TOPIC t_id="D0804">
       <H h_id="699">
            <H_sentence>People were forced to leave their pets behind when they evacuated New
              Orleans.</H_sentence>
            <text doc_id="AFP_ENG_20050907.0111" s_id="11" evaluation="YES">Thousands of people were forced
              to leave their pets behind when they evacuated New Orleans.</text>
          …
         </H>
         …
         <H h_id="702">
            <H_sentence>On Friday, September 16, 2005, an animal welfare group warned that time was running out for
            pets abandoned in the wake of Hurricane Katrina.</H_sentence>
         </H>
          …
    </TOPIC>
    <TOPIC t_id="D0805">
          …
    </TOPIC>
        …
  </entailment_corpus>
```

Note that for those H’s which do not have entailing sentences (e.g. h_id=”702” in the example above), the <H> element is given, but it does not contain any <text> elements.

<u>Example 2: Item A</u>

```
<HYPOTHESES>
    <H h_id="699">
    <text> People were forced to leave their pets behind when they evacuated New Orleans. </text>
    <ref doc_id="AFP_ENG_20050907.0111" s_id="11">Thousands of people were forced to leave their pets behind
    when they evacuated New Orleans.</ref>
    </H>
    …
</HYPOTHESES>
```

Example 3: Item B

```
<topic id="D0804">
    <H h_id="699">
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="0" lucene_score=" 0.89145654" />
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="4" lucene_score="0.28941703" />
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="5" lucene_score="0.19954799" />
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="6" lucene_score="0.14987582" />
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="9" lucene_score="0.14966099" />
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="11" lucene_score="0.13179718"/>
        <CANDIDATE doc_id="AFP_ENG_20050907.0111" s_id="14" lucene_score="0.11532254" />


            …
    </H>
            …
</topic>
```

Furthermore, also Cluster B documents, from which some hypotheses are taken, are given, grouped together in a single file. This item is not directly relevant to the task but is provided for documentation purposes only.

Test Set

The Test Set will include the same items as the Development Set except the gold standard, which will not be distributed, as it will be used for the evaluation of the system performances.

Participants are reminded that the Test Set is blind, and must not be analyzed before submitting the results.

Submission Format

Participants are allowed to submit up to 3 runs. For each run, a single file containing all the candidate sentences for which the entailment decision is "YES" must be submitted. The format is the same as the gold standard released with the Development Set, except that only IDs are required, e.g.:

```
<entailment_corpus>
    <TOPIC t_id="804">
        <H h_id="699">
        <text doc_id=" AFP_ENG_20050907.0111" s_id="11"/>
        <text doc_id=" AFP_ENG_20050907.0111" s_id="34"/>
        ...
        </H>
        <H h_id="700">
        </H>
        ...
        </TOPIC>
        <TOPIC t_id="D0805">
        …
    </TOPIC>
    …
</entailment_corpus>
```

Note that, if no entailing sentences are found for a given H (e.g. h_id="4" in the example above), the <H> element for that given hypothesis must be returned anyway, but with no <text> element. Similarly, if a system cannot find any entailing sentences for an entire topic, the related <TOPIC> element must contain all the <H> elements, with no <text> elements.

At the time of submission, each team will be asked to fill out a submission form at the RTE-7 website, stating a number (1-3) for the run, used to differentiate between the team's runs for the task.

3.4 Result Evaluation

System results will be compared to the human-annotated gold standard and the metrics used to evaluate system performances will be Micro-Averaged Precision, Recall, and F-measure.

4. NOVELTY DETECTION SUBTASK

4.1 Task description

The Novelty Detection subtask is based on the Main task and is aimed at specifically addressing the interests of the Summarization community, in particular with regard to the Update Summarization task, focusing on detection of novelty in Cluster B documents.
The task consists of judging if the information contained in each H (drawn from the cluster B documents) is novel with respect to the information contained in the set of Cluster A candidate entailing sentences. If for a given H one or more entailing sentences are found, it means that the content of the H is not new. On the contrary, if no entailing sentences are detected, it means that the information contained in the H is regarded as novel.

The Novelty Detection task requires the same output format as the Main task – i.e. no additional type of decision is needed[6]. Nevertheless, the Novelty Detection task differs from the Main task in the following ways:

1) The set of H's is not the same as that of the Main task (see 4.2);
2) The system outputs are scored differently, using specific scoring metrics designed for assessing novelty detection (see 4.4).

Participants in this task have the opportunity to tune their systems specifically for novelty detection. Within this setting, it is particularly relevant that each H is processed and judged for entailment independently from the other H's.

4.2 Data set description

The Novelty Detection data set is similar to the Main task data set, described in Section 3.3, except that it contains a different set of H's and corresponding candidate sentences. In the Main task, H's are taken both from Cluster B automatic summaries and Cluster A sentences (see footnote 4). However, the H's that are not taken from the automatic summaries are less interesting from a Summarization perspective, because they have relatively numerous entailing sentences in the Cluster A corpus and can be more easily recognized as non-novel by the summarization systems. Therefore, the Novelty Detection data contain all and only the H's taken from Cluster B automatic summaries, which reflect more directly the output of actual summarization systems.[7]

4.3 Data set and submission format

As the Novelty Detection data set is a subset of the Main task data set, the format description is the same as provided in Section 3.3 for the Main task.

As for the Main task, the following items will be distributed also for the Novelty Detection task:

1) the Development Set gold standard specific for the Novelty Detection task;

[6] This means that, as in the Main task, when a H judged as novel, i.e. no entailing sentences for that H are found among the Cluster A candidate entailing sentences, no text must be returned; meanwhile when a H is judged as containing non-novel information, all the entailing sentences must be returned as justification of the non-novelty judgment.

[7] The Novelty Detection H set is made up of all the H's needed to represent the prominent majority of the content of the automatic summaries of the three best Summarization systems. A typical kind of information which is not covered by the H's are the sources of news (e.g., given the sentence "*Norwegian Foreign Minister will visit Sri Lanka, the Norwegian Embassy said in a press release.*", the piece of information "*the Norwegian Embassy said in a press release"* will not be included in the H set).

2) for each topic:

– *Item A*: a list of hypotheses (see Example 2). The element <ref> contains the corpus sentence from which the H is taken. The sentence itself is given for documentation purposes only and must not be used as additional data to deduce entailment. Its attribute “doc_id” indicates the ID of the document in which the sentence is found. This attribute encodes the publication date of the document and can be used to time anchor the hypothesis (e.g. *AFP_ENG_20050728.0294*, where *20050728* is the publication date in the format *YYYYMMDD*). Note that in the Novelty Detection task, the time of H in is always later than the time of T, due to the fact that H’s are taken from Cluster B, made up of more recent documents.

– *Item B*: for each H listed in Item A, the list of the ID numbers of Cluster A candidate sentences to be judged for entailment. Note that for each candidate sentence its Lucene ranking score is given.

– *Item C*: the set of Cluster A documents for that topic.

SUBMISSION FORMAT

Participants can submit up to 3 additional runs specifically for the Novelty Detection task. The output format required is the same as for the Main task (see 3.3). No additional information is needed, as the novelty detection decision is derived automatically from the number of entailing sentences provided for each H (0 or more).

4.4 RESULT EVALUATION

As in the Main task, the system results will be compared to the human-annotated gold standard; however, as mentioned above, only the H’s taken from the automatic summaries will be evaluated for novelty detection.

Two scores will be used to evaluate the system performances on the Novelty Detection task:

1) The primary score will be Precision, Recall and F-measure computed on the binary novel/non-novel decision. The novelty detection decision is derived automatically from the number of justifications provided by the system (i.e. the entailing sentences retrieved for each H) - where 0 implies ‘novel’, 1 or more ‘non-novel’.

2) The secondary score will measure the quality of the justifications provided for non-novel H’s, that is the set of all the sentences extracted as entailing the H’s. The metrics used to this purpose will be Micro-averaged Precision, Recall and F-measure.

5. ABLATION TESTS

Like in RTE-6, ablation tests are required for systems participating in the RTE-7 Main task, in order to collect data to better understand the impact of both knowledge resources and tools used by RTE systems and evaluate the contribution of each resource to systems' performance.

Please remember that ablation tests are mandatory for participation in the RTE-7 Main task, and must be performed by all participants. If it is impossible to carry out any tests due to the system architecture, this must be explicitly declared (see instructions below, point 6).

An ablation test consists of removing one module from a complete system, and rerunning the system on the test set with the other modules (excluding the module being tested). Comparing the results to those obtained by the complete system, it is possible to assess the practical contribution given by the individual module.

The instructions on how to perform the ablation tests and submit the results are as follows:

1. The output of the ablation tests will be submitted through an online submission form (TBA) at the RTE-7 website.
2. The ablation tests must be carried out on one or more of the runs submitted to the RTE-7 Main task.
3. For each ablation test, a single knowledge resource or tool must be *added to* or *removed from* the base system that produced the selected run in the Main task, and the system must be rerun; the output must have the same format as the one obtained by running the base system. Tests where resources or tools are added rather than removed, can be considered ablation tests, as only one knowledge resource or tool is evaluated, with the difference that the system that omits the ablated resource/tool is the system participating in the Main task.
4. An ablation test must be carried out on each of the knowledge resources and tools used by the base system in the Main task. If too many knowledge resources and/or tools are used and it is infeasible to carry out tests on all of them, at least three of the knowledge resources and/or tools must be tested, specifically those which are thought to have the greatest impact on the overall performance of the complete system. *In case a system does not use any resources or tools, participants are encouraged to submit at least one ablation test where one resource or tool is added to the base system.*
5. Each ablation test run will be accompanied by a description indicating the resource or tool which has been ablated; if the ablation test requires modifying the system, this should also be explained in the description.
6. If ablation tests cannot be done due to system architecture, this should be explained in the ablation test submission form.

5. PROPOSED SCHEDULING

– April 29:	Release of Development Set
– August 29:	Main task: Release of Test Set
– September 8:	Main task: Deadline for task submissions
– September 15:	Main task: Release of individual evaluated results
– September 29:	Main task: Deadline for ablation tests submissions
– October 6:	Main task: Release of individual ablation test results
– October 25:	Deadline for system reports (workshop notebook version)
– November 14-15:	TAC 2011 workshop in Gaithersburg, Maryland, USA

REFERENCES

Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Lo Leggio, M., Magnini, B. (2009a). Considering Discourse References in Textual Entailment Annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa, Italy.

Bentivogli, L., Dagan, I., Dang H.T., Giampiccolo, D., Magnini, B. (2009b). The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the TAC 2009*, Gaithersburg, MD, USA.

Bentivogli, L., Clark, P., Dagan, I., Dang H.T., Giampiccolo, D. (2010). The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the TAC 2010*, Gaithersburg, MD, USA.

Dagan, I., Glickman, O., Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), *Machine Learning Challenges*. Lecture Notes in Computer Science, Vol. 3944, Springer.

Appendix A: Annotation Guidelines

In Recognizing Textual Entailment within a corpus, entailment judgment should simulate the inferences that would be made by a reader of a text unit T, based on that particular text and on the collection to which the text belongs. We assume that when reading T the reader performs two processes, namely:

A) s/he fully interprets its meaning, as intended to be communicated by the writer

B) optionally, s/he makes additional inferences from T, beyond the explicit intended scope of its meaning, if needed to infer H.

To perform these processes, the reader incorporates prior knowledge (i.e. knowledge assumed available to the reader prior to reading the sentence) of the following types:

1) Knowledge about all explicit and implied references within the sentence, which are part of its intended meaning and are thus needed for its proper interpretation. T's may contain various types of references, such as references to persons, locations, dates, events, actions. Examples of these different kinds of references are given below:

- [***H1***: *2003 UB313 is bigger than Pluto*]
  T1: *"It's definitely bigger than Pluto", he said of the body made up of ice and rock.*
  Reference knowledge: It = the body = 2003 UB313 (a "planet" code name); he = Michael Brown

- [***H2***: *Russia requested international help to rescue the AS-28*]
  T2: *Russia resisted international assistance in that crisis, and made a series of false statements about its problems at sea.*
  Reference knowledge: that crisis= Kursk's sinking (previous sentence: Russia's last prominent submarine crisis, in 2000, when the nuclear submarine Kursk sank after on-board explosions in shallow water in the Barents Sea).

The use of reference information does not cover just references to prior sentences in the same text, but also reference information which is globally available from the corpus, for example:

- *[**H3**: High temperatures shrink the Arctic ice]*
  T3: *Rising air and ocean temperatures have been cited.*
  Reference knowledge: Time= present (2005); Loc= Arctic; have been cited= as a cause for ice shrinking. While the problem of ice shrinking is mentioned in the previous sentence, neither time nor location are explicitly mentioned in the document.
- *[**H4**:About 50 people were killed in the attack]*
  T4: *Forty-eight people died.*
  Reference knowledge: died -> in the attack (previous sentence); attack -> Loc=London & Time=July 7, 2005.

2) Language knowledge needed to fully interpret the sentence meaning, for example:

- *[**H5**: Mine accidents cause deaths in China]*
  T5**:** *So far this week, four mine disasters have claimed the lives of at least 60 workers and left 26 others missing.*
  Language knowledge: to claim the lives = to cause the death of some people

3) Common background world knowledge, which can be needed both to interpret the sentence (process A), and to make additional inferences in order to infer H (process B). For example:

- ***H6:*** *The ice is melting in the Arctic.*

T6: *The scene at the receding edge of the Exit Glacier in Kenai Fjords National Park in Alaska was part festive gathering, part nature tour with an apocalyptic edge.*
World knowledge**:** Alaska is in the Arctic; the edge of the glacier is receding because the ice is melting.

We say that T entails H only if for certain *knowledge* that can be incorporated to interpret T and to infer H (Processes A+B)

TEXT AND KNOWLEDGE ENTAIL H, **BUT** KNOWLEDGE ALONE CANNOT NOT ENTAIL H

This means that the knowledge which should be incorporated to allow inferring H from T should not entail H alone. In other words, it is not allowed to validate H's truth regardless of T.

For example:

- ***H7:*** *Ice is melting in the Arctic*
  T7: *Global warming causes permafrost's shrinking.*
  Reference knowledge: TIME= present (2005); LOC=Arctic
  Language knowledge: permafrost = a permanently frozen soil at variable depth below the surface in frigid regions of a planet - as earth
  Background world knowledge: global warming = increase in the average temperatures; high temperatures melt the ice; the permafrost is shrinking because the ice contained in it is melting
  Entailment judgement: H entailed by T (permafrost's shrinking=ice is melting) + knowledge
  ENTAILMENT: YES

- ***H8:*** *The ice is melting in the Antarctic.*
  T8: *"This time, the problem is man-made and if we don't take steps, the damage will be worse," he said.*
  Reference knowledge**:** the problem = the ice melting in the Antarctic; he=John Barry; this time=present (2005)
  Entailment judgement: H entailed by knowledge alone
  ENTAILMENT: NO ENTAILMENT

As it can be seen in the above examples, T7 contributes the needed information to infer H (by saying that permafrost is shrinking), which may not be inferred without that information, while T8 does not contribute any information needed for the entailment judgment.

It is important to note that knowledge about the topic which has been acquired from other sentences in the corpus, but which is not part of T's interpretation (via some reference), cannot be used in order to establish entailment from T. This applies in particular to cases where a Text makes an assertion about a future event, but a priori world knowledge suggests that the event cannot be predicted with high probability.

For example:

- ***H9:*** *Orahn Pamuk went to court on 16 December 2005*
  T9*: He said "The trial is expected to start on December 16".*
  Topic knowledge (not allowed): the trial indeed started on December 16 (from another document)
  Entailment judgment: Even if we know from other documents of the corpus that H9 is true (that is, that the trial indeed started when expected), this knowledge cannot be used in the entailment process.
  ENTAILMENT: NO ENTAILMENT

In the example above, a priori knowledge suggests that trial dates sometimes (not infrequently) change from their expected dates, so the actual trial date in H9 cannot be inferred from T9. Opinions may differ on the predictability of certain events, but in any case, the a priori world knowledge about the predictability of an event should guide the entailment decision, rather than specific knowledge that is learned from other sentences in the document set.

Other examples of entailment judgment

H: *2003 UB313 is larger than Pluto.*
T*: "It's definitely bigger than Pluto", he said of the body made up of ice and rock.*
Reference knowledge: It = the body = 2003 UB313
Language knowledge: bigger = larger
ENTAILMENT: YES

H: *Russia requested international help to rescue the AS-28.*
T: *Russia resisted international assistance in that crisis, and made a series of false statements about its problems at sea.*
Reference knowledge: that crisis= Kursk's sinking (previous sentence: Russia's last prominent submarine crisis, in 2000, when the nuclear submarine Kursk sank after on-board explosions in shallow water in the Barents Sea) =/= AS-28 accident.
ENTAILMENT: NO

H*: Airbus A380 can carry 540 people*
T*: This plane can carry 600 people.*
Reference knowledge: This plane = Boeing 747
Entailment judgment: The entity mentioned in T is not the entity mentioned in H; the plane mentioned in T, able to carry 600 people, is a Boeing 747 which is explicitly mentioned in the previous sentence of the document, where it is compared to Airbus A380.
ENTAILMENT*:* NO

H: *High temperatures shrink the Arctic ice.*
T: *Rising air and ocean temperatures have been cited.*
Reference knowledge: Loc= Arctic; have been cited= as a cause for ice shrinking (previous sentence).
Language knowledge: to rise = to increase
Background world knowledge: if temperatures increase, they become higher
Entailment judgement: H is entailed by T (rising temperatures) + knowledge
ENTAILMENT: YES

H:*The ice is melting in the Arctic*
T: *Rising air and ocean temperatures have been cited.*
Reference knowledge: Loc= Arctic; have been cited= as a cause for ice shrinking (previous sentence).
Entailment judgement: all the information necessary to infer H is contained in the reference knowledge "as a cause for ice shrinking"
ENTAILMENT: NO

H: *About 50 people were killed in the attack.*
T: *Forty eight people died.*
Reference knowledge: died -> in the attack (previous sentence).
Language knowledge: *kill* = to make a person or animal *die*
Entailment judgement: H is entailed by T (48 people died) + knowledge (in the attack)

ENTAILMENT: YES

H: *Michael Brown discovered 2003 UB313*
T: *Then on Jan. 5 -- not Jan. 8 as he had said at his news conference -- he finally found one he could call Xena.*
Reference knowledge: he = Brown; one = Xena = 2003 UB313.
Entailment judgement: H entailed by T (he found one) + knowledge
ENTAILMENT: YES

H: *Mine accidents cause deaths in China.*
T**:** *So far this week, four mine disasters have claimed the lives of at least 60 workers and left 26 others missing.*
Reference knowledge: LOC= in China.
Language knowledge: to claim the lives = to cause the death of some people; disaster = accident.
Entailment judgement: H entailed by T + knowledge
ENTAILMENT: YES

H: *Ice shelves are thinning*
T: *"Ice is thinning at the rate of tens of meters per year" on the peninsula, with glacier elevations in some places having dropped by as much as 124 feet in six months, it found.*
Entailment judgement: T does not specifically refer to ice shelves
ENTAILMENT= NO

H: *The ice is melting in the Arctic.*
T: *The scene at the receding edge of the Exit Glacier in Kenai Fjords National Park in Alaska was part festive gathering, part nature tour with an apocalyptic edge.*
Background world knowledge**:** Alaska is in the Arctic; the edge of the glacier is receding because the ice is melting.
Entailment judgement: if the Exit Glacier is receding, then the ice is melting in the Arctic.
ENTAILMENT: YES

H: *2003 UB313 has a moon.*
T: *The moon was first spotted by a 10-meter telescope at the W.M. Keck Observatory in Hawaii on Sept. 10.*
Reference knowledge: The moon = 2003 UB313's moon
Entailment judgment = T is entailed by Knowledge alone
ENTAILMENT= NO

H: *The Airbus A380 flew its maiden test flight.*
T: *Its second flight was absolutely successful.*
Reference knowledge**:** Its = of Airbus A380
Background world knowledge: If the airbus flew its second flight, then it flew also a first flight (the maiden test flight)
Entailment judgement: H is entailed by T (A380 second flight) + knowledge (if second flight then first flight)
ENTAILMENT = YES.

H: *The AS-28 mini-submarine was trapped underwater*
T: *In televised comments, Pacific Fleet spokesman Capt. Alexander Kosolapov said there was contact with the sailors, who were not hurt, and that authorities were preparing to send a similar vessel to assess the situation.*

Reference knowledge**:** the sailors = of the mini-submarine trapped underwater; a similar vessel=a submarine
Entailment judgement: H entailed by knowledge alone
ENTAILMENT: NO

H*: An attack occurred in London*
T*: She first spoke admiringly of how Londoners handled themselves during the attacks and immediately after, remaining remarkably calm during a morning that could have devolved into chaos but remained far from it.*
Reference knowledge**:** attacks = the 4 attacks composing the bomb attack
Language knowledge**:** Londoners = inhabitants of London
ENTAILMENT: YES

H: *The European Union was concerned about freedom of expression in Turkey.*
T: *Rehn said the new penal code "does not provide sufficient protection for the freedom of expression" and the Turkish government should "close the loopholes in the code."*
Reference knowledge**:** Rehn = Olli Rehn = EU Enlargement Commissioner
Background world/language knowledge: EU Enlargement Commissioner = EU (metonymy)
Entailment judgement: Rehn is considered a reference to the EU.
ENTAILMENT: YES

H: *Ice is melting in the Antarctic*
Ta: *The glaciers are retreating in the Antarctic.*
Tb: 9000 years ago *the ice shelves in the Antarctic melted down almost completely.*
Time Reference: TIME in H = present (use of present continuous).
TIME in Ta = Present, 2005 (use of present continuous).
TIME in Tb = Past (“9000 years ago”).
Language knowledge: Glaciers are made of ice. Ice shelves are made of ice.
Background world knowledge: Glaciers retreat because ice melts. Ice shelves melts because ice melts.
Entailment judgement a: The time of T coincide with the time of H.
Entailment judgement b: The time of T does not coincide with the time of H.
ENTAILMENT of Ta**=** YES
ENTAILMENT of Tb**=** NO