5TH TEXTUAL ENTAILMENT CHALLENGE @ TAC 2009

TEXTUAL ENTAILMENT SEARCH PILOT

Task Guidelines

Version March 16, 2009

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

1. INTRODUCTION

During the RTE4 Challenge workshop at TAC 2008 the need to move towards more realistic scenarios was stressed once again, both by organizers and participants. In the campaigns held so far, the necessity of getting acquainted with the textual entailment task had motivated proposing only test sets in which the pairs were artificially adapted in order to facilitate the study of the different entailment phenomena; however, the progress that has been made, now allows to make a step forward and start to test RTE systems against real data.

In order to meet this demand, a Textual Entailment Search Pilot task has been set up, aimed principally at:
1) producing 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;
2) analyzing the potential impact of textual entailment recognition on a real NLP application task, namely the Summarization task as proposed by the Summarization community in the 2008 Text Analysis Conference at NIST.

This document provides a definition of the Pilot Search Task and a description of the data set, together with the instructions on how to take part in the exercise.

2. TASK DESCRIPTION

The Textual Entailment Search task consists in finding all the sentences in a set of documents that entail a given Hypothesis.

The task is situated in the Summarization application setting, where the Hypothesis (H) is taken from a Summary Content Unit[1] (SCU), and the systems must find all the entailing sentences (Ts) in a corpus of 10 newswire documents about a common topic.

We assume the standard definition of Textual Entailment [Dagan et al., 2006] as a directional relationship between two text fragments, which we term the Text (T) and the Hypothesis (H). We say that:

[1] SCUs are sub-sentential content units, not bigger than a clause, taken from a corpus of manually-made summaries. SCUs are used in the evaluation of Summarization tasks

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 we require 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 example below presents a hypothesis referring to a given topic and the entailing sentences found in the corpus of 10 documents about the same topic:

<H_sentence>Seven submariners were onboard the AS-28.</H_sentence>
<text doc_id="AFP_ENG_20050804.0725" s_id="1" evaluation="YES">The Russian military was racing against time early Friday to rescue a small submarine that had become trapped on the seabed with seven crew aboard, the Ria-Novosti news agency reported.</text>
<text doc_id="APW_ENG_20050807.0129" s_id="2" evaluation="YES">All seven aboard the AS-28 mini-submarine appeared to be in satisfactory condition, naval spokesman Capt. Igor Dygalo said.</text>
<text doc_id="APW_ENG_20050807.0129" s_id="8" evaluation="YES">It was carrying six sailors and a representative of the company that manufactured it.</text>
<text doc_id="NYT_ENG_20050805.0181" s_id="2" evaluation="YES">The seven men on board were said to have as little as 24 hours of air.</text>
<text doc_id="AFP_ENG_20050805.0571" s_id="6" evaluation="YES">There are seven crew members aboard the vessel, stranded on the ocean floor at a depth of around 190 meters (623 feet) in a bay off the coast of the Kamchatka peninsula in Russia's Far East region.</text>

As it can be seen from the above example, the major difference with respect to the main task is that in the framework of the main exercise, where isolated T/H pairs are given, both Text and Hypothesis are artificially created in a way that they do not contain references to information outside the H/T pair, and hence the context necessary to judge the entailment relation is given by T. Only language and world knowledge are needed within the main task, while reference knowledge is typically not required.

In contrast, in the Entailment Search task both Text and Hypothesis 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.

As specifically regards the reference to the time in which a sentence was written - which impacts particularly the interpretation of tenses - it must be taken into account that while T (a sentence in a document) is naturally anchored to the publication date of the document containing it, H is anchored (by construction) to the time at which a summary of all the topic documents was written. The anchor time of the hypotheses related to a given topic can be conventionally fixed as the day after the publication of the last topic document. See, for instance, the following example:

<H_sentence> European Union officials were in Turkey to observe the trial against Orhan Pamuk.</H_sentence>
<text doc_id="APW_ENG_20051215.0108" s_id="12" evaluation="YES"> *EU officials are now in Turkey to observe Pamuk's trial*.</text>

In order to clarify the process necessary to interpret the entailment relation in the Search task, Appendix A presents the guidelines which have been followed by the annotators when creating the data set.

The Entailment Search Task requires the retrieval of entailing sentences only, hence contradicting sentences are not to be taken into account, and the entailment judgment may be seen as a two-way decision between "yes" and "no" entailment.

3. DATA SET DESCRIPTION

The Entailment Search Data Set is based on the data created for the TAC 2008 and 2009 Update Summarization tasks. More precisely, the Development Set is composed of 10 topics randomly chosen from the 48 topics of the 2008 exercise, whereas the Test Set will be composed of 10 topics that will be randomly taken from the TAC 2009 SUM Update data.

For each topic, the Textual Entailment Search data consist of:

1. Between 6 and 10 Hypotheses referring to the topic.
Hypotheses are created on the basis of the pyramid SCUs, which are taken from 4 manually-made summaries of the documents.

2. A set of 10 documents.
Two clusters were available from the SUM Update data: Cluster A which is made up of the first 10 texts in chronological order of the whole SUM cluster, and Cluster B which is made up of the last 10 texts. The 10 documents of the A cluster were chosen because in the related manual summaries there is no assumption of previous knowledge about the topic.

Since the sentence is the most relevant unit for the Summarization task, all documents have been manually split into sentences, which represent the Texts to be judged for entailment.

As far as entailment is concerned, it must be taken into account that:

a) Contradictions are not considered in this Pilot task, and thus the entailment judgment choice must be between "yes" and "no" entailment.
b) For each hypothesis there is at least one entailing sentence in the corpus.
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 time of publication, (ii) a headline, and sometimes (iii) an explicit dateline.
   - DOC: id="AFP_ENG_20051216.0104" type="story"
     HEADLINE: Turkish court meets in Pamuk case
     DATELINE: ISTANBUL, Dec 16
e) The headline sentence must be judged for entailment as well[2].
f) Documents may contain non-informative sentences which must be judged as ***NO*** entailment
   – EX: *"_____", "On the net:", "1994", "New York:", "(Begin optional trim)"*

[2] In the document file, the headline is reported twice: at the beginning of the document as in the original SUM file, and as the first sentence (S_id="0") to be annotated.

In order to download the Search task data (topic sample, development set, test set) it is necessary to submit to TAC 2009 the following user agreements:
1. Agreement Concerning Dissemination of TAC Results
2. AQUAINT-2 Agreement

A link to all the user agreements can be found at the TAC 2009 home page (http://www.nist.gov/tac/2009/).

4. DATASET AND SUBMISSION FORMAT

The following items will be distributed as Development Set:
- a list of hypotheses for each topic
- a set of documents for each topic
- a gold standard, which consists of a single file in the following XML format:

```
<entailment_corpus>
  <TOPIC t_id="D0801-A">
    <H h_id="54">
       <H_sentence> Airbus A380 flew its maiden test flight.</H_sentence>
       <text doc_id="NYT_ENG_20050427.0044" s_id="1" evaluation="YES">With a whisper
    more than a roar, the largest passenger airliner ever built, the Airbus 380, took off on its maiden
    flight Wednesday.</text>
       …
    </H>
    <H h_id="55">
       …
    </H>
  </TOPIC>
  <TOPIC t_id="D0802-A">
    …
  </TOPIC>
    …
</entailment_corpus>
```

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

Participants are required to submit a single file containing all the sentences for which the entailment decision is “YES”. The format is the same as the gold standard released with the Development Set, except that only IDs are required:

```
<entailment_corpus>
  <TOPIC t_id="D0801-A">
    <H h_id="54">
       <text doc_id="NYT_ENG_20050427.0044" s_id="1"/>
       <text doc_id=" APW_ENG_20050806.0726" s_id="7"/>
       …
    </H>
    <H h_id="55">
       …
```

```
    </H>
      ...
  </TOPIC>
  <TOPIC t_id="D0802-A">
    ...
  </TOPIC>
    ...
</entailment_corpus>
```

5. RESULT EVALUATION

System results will be compared to a human-annotated gold standard and the metrics used to evaluate system performances will be Precision, Recall, and F-measure. Both micro averaged and macro averaged results will be made available to participants.

6. SCHEDULING

- April 3: Release of Development Set
- September 2: Release of Test Set
- September 9: Deadline for Task submission
- September 18: Release of individual evaluated results
- Mid-October: Deadline for systems' reports

REFERENCES

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.), *Machine Learning Challenges*. Lecture Notes in Computer Science, Vol. 3944, pp. 177-190, Springer, 2006.

Appendix A: Annotation Guidelines

Within the Textual Entailment Search Task, 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 the text belongs to. 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. Ts 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 specific knowledge about the topic which has been acquired from other sentences in the corpus, but which is not part of T's interpretation, cannot be used in the entailment process. This applies in particular to cases where T makes an assertion about a future event, but common background knowledge suggests that the event cannot be predicted with high probability.

- ***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, common background 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 (and common background knowledge about the predictability of an event should not be changed based on 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 = 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.a*=** YES
***ENTAILMENT.b*=** NO