6TH TEXTUAL ENTAILMENT CHALLENGE @ TAC 2010

KNOWLEDGE BASE POPULATION VALIDATION PILOT

Task Guidelines

1. INTRODUCTION

During the TAC 2009 workshop, the importance of sharing a common framework among the three TAC tracks was reiterated, and collaboration among these text understanding fields was further strengthened.

The successful experience of the RTE-5 Pilot task, situated in the Summarization scenario, not only indicated the need for a deeper investigation of the potential impact of Textual Entailment on the Summarization field, which is being carried out within the RTE-6 Main task, but also pointed out opportunities to broaden the interaction between RTE and other application areas at TAC.

Aiming to more directly address the needs of NLP applications, a Knowledge Base Population (KBP) *Validation* Task is proposed as a Pilot within RTE-6. This task is based on the TAC KBP *Slot Filling* task (McNamee and Dang, 2009), and is meant to show the potential utility of RTE systems for Knowledge Base Population.

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

2. TASK DESCRIPTION

The KBP Validation Pilot is situated in the Knowledge Base Population scenario, and aims at validating the output of the systems participating in the KBP Slot Filling task by using Textual Entailment techniques.

The KBP *Slot Filling* task is focused on searching a collection of newswire and Web documents and extracting values for a pre-defined set of attributes (a.k.a. "slots") for target entities. Given an entity in a knowledge base and an attribute for that entity, systems must find in a large corpus the correct value(s) for that attribute and return the extracted information together with a corpus document supporting it as a correct slot filler.

The RTE KBP *Validation Pilot* is based on the assumption that an extracted slot filler is correct if and only if the supporting document entails an hypothesis created on the basis of the slot filler[1].

For example, consider the following slot filler and supporting document returned by a KBP system for the "residences" attribute for the target entity "Chris Simcox":

[1] Another example of the use of Textual Entailment techniques to validate the output of real application systems is represented by the Answer Validation Exercise, which was proposed within the Question Answering track at the CLEF Campaign from 2006 to 2008 (Peñas et al., 2007).

Slot Filler: “Tucson, Ariz.”
Document ID: NYT_ENG_20050919.0130.LDC2007T07

If the slot filler is correct, then the document NYT_ENG_20050919.0130.LDC2007T07 must entail one or more of the following Hypotheses:

H1: Chris Simcox lives in Tucson, Ariz.
H2: Chris Simcox has residence in Tucson, Ariz.
H3: Tucson, Ariz. is the place of residence of Chris Simcox
H4: Chris Simcox resides in Tucson, Ariz.
H5: Chris Simcox’s home is in Tucson, Ariz.

The KBP Validation Task consists of determining whether a candidate slot filler is supported in the associated document. Each slot filler that is submitted by a system participating in the KBP *Slot Filling* task creates one evaluation item (i.e. a T-H “pair”) for the RTE-KBP Validation Pilot, where T is the source document that was cited as supporting the slot filler, and H is a set of simple, synonymous Hypotheses created from the slot filler.

A distinguishing feature of the KBP Validation Pilot is that the resulting T-H pairs differ from the traditional pairs. In particular:

a. T is an entire document (vs. single sentences or paragraphs)
b. H is not a single sentence but a set of roughly synonymous sentences representing different linguistic realizations of the same slot filler.

Another major characteristic of the KBP Validation task, which distinguishes it from the other RTE challenges proposed so far, is that the RTE data set is created *automatically* from KBP *Slot Filling* participants’ submissions, and the gold standard annotations are automatically derived from the KBP assessments. Moreover, as H’s are created automatically, they can be ungrammatical.

3. DATA SET DESCRIPTION

The RTE-6 KBP Validation data set is based on the data created for the KBP 2009 and 2010 *Slot Filling* Task. More precisely, the Development Set was created from the KBP slot-filling system output and slot-filler assessments from KBP 2009, whereas the Test Set will be created from corresponding data from KBP 2010. The definitions of the slots and guidelines for assessing slot fillers for KBP 2009 and 2010 are available on the TAC 2010 KBP web site at http://nlp.cs.qc.cuny.edu/kbp/2010/annotation.html.

The creation of the RTE-6 Pilot task data set is semi-automatic and takes as starting point (i) the extracted slot-fillers from multiple systems participating in the KBP *Slot Filling* task and (ii) their assessments[2].

A first manual phase, preliminary to the automatic generation of the H’s of the data set, requires that several “seed” linguistic realizations of templates are created for each target slot, expressing the relationship between the target entity and the extracted slot filler. For example, given the attribute “origin” belonging to a target entity of type “person”, the following templates are manually created:

[2] The Slot Filling task can be viewed as a more traditional Information Extraction task. The methodology used for creating the T-H pairs in this Pilot was already adopted for the (manual) creation of those pairs in the Main Task data sets from RTE-1 to RTE-5, which represented the IE application setting. In order to create those IE pairs, hypotheses were taken from the relations tested in the ACE tasks, while texts were extracted from the outputs of actual IE systems, which were fed with relevant news articles. Correctly extracted instances were used to generate positive examples, and incorrect instances to generate negative examples.

Template 1: ***X'**s* `origins are in` ***Y***
Template 2: ***X*** `comes from` ***Y***
Template 3: ***X*** `is from` ***Y***
Template 4: ***X*** `origins are` ***Y***
Template 5: ***X*** `has` ***Y*** `origins`
Template 6: ***X*** `is of` ***Y*** `origin`

Then, each slot filler submitted by a system participating in the KBP *Slot Filling* task represents one evaluation item and is used to automatically create an RTE T-H pair. The T corresponds to the corpus document supporting the answer (as identified by the KBP system), while the H is created by instantiating all the templates for the given slot both with the name of the target entity (X) and the slot filler extracted by the system (Y). Providing all the instantiated templates of the corresponding slot for each system answer has the consequence that each T-H pair does not contain only one single H, but a set of synonymous H's. This setting has the property that for each example either all H's for the slot are entailed or all of them are not.

Moreover, it is important to note that, due to the way that H's are created, systems must be prepared to deal with ungrammatical H's. In fact, while the H's templates are fixed, the slot fillers returned by the systems are strings which can be incomplete, include extraneous text, or belong to a POS which is not compatible with that required by a specific H template. In the following example, given (i) the H templates for the slot "origin", (ii) the target person entity "Chris Simcox" and (iii) a correct slot filler "Canadian", we obtain both grammatical and ungrammatical H's within the same evaluation item:

H1: Chris Simcox's origins are in Canadian
H2: Chris Simcox comes from Canadian
H3: Chris Simcox is from Canadian
H4: Chris Simcox origins are Canadian
H5: Chris Simcox has Canadian origins
H6: Chris Simcox is of Canadian origin

The RTE gold standard annotations are automatically derived from the KBP assessments, converting them into Textual Entailment values. The assumption behind this process is that the KBP judgment of whether a given slot filler is correct coincides with the RTE judgment of whether the text entails the template instantiated with the target entity and the automatically extracted slot filler[3]. Note that contradictions are not considered in this task, and thus the entailment judgment is either "YES" or "NO" entailment.

Moreover, because no temporal qualifications are defined for the KBP slots, differences in verb tense between the Hypothesis and Document Text in the RTE KBP Validation task *must be ignored*. In the KBP Slot Filling task, for example, "Tucson, Ariz." is considered a correct slot filler for the "residence" attribute of the target entity "Chris Simcox" if the supporting document contains the text "Chris Simcox lived in Tucson, Ariz., before relocating to Phoenix"; therefore, in the KBP Validation task, the Hypothesis "Chris Simcox lives in Tucson, Ariz." must be judged to be entailed by the same document.

[3] The KBP assessments were not binary values and, therefore, a mapping was necessary to convert KBP assessments into entailment values: "correct" and "redundant" KBP judgments map into YES entailment; "wrong" judgments map into NO entailment; "inexact" judgments can result both in YES and NO entailment values, and for this reason RTE pairs involving "inexact" KBP judgments have been excluded from the data set. Removing inexact pairs and other pairs that were deemed unsuitable for the KBP Validation task resulted in a Development Set of 9,462 T-H pairs. Some nuanced cases remain where the KBP judgment and the RTE judgment may not coincide, but such cases are expected to be rare and have minuscule impact on the KBP Validation results.

As the KBP Validation Test Set will be created from the KBP 2010 data, it is important to note that the Test Set and Development Set will possibly differ in some aspects:

- the size of the data set and the ratio between positive and negative pairs may vary, depending on the number of KBP 2010 systems' submissions and on their performances.
- the proportion of web documents with respect to newswire documents may vary, as the 2010 KBP corpus will contain a higher number of web documents.
- some location slots ("place of birth", "place of death", "residence", and "headquarters") in the KBP 2009 data have been expanded into separate, more specific slots (e.g., "city of birth", "state or province of birth", and "country of birth"); therefore, the templates may be changed to reflect the more specific semantics of the slot.

The RTE-6 KBP Validation data (Development Set and Test Set), are distributed by the Linguistic Data Consortium (LDC). Registered RTE-6 teams may request KBP Validation data from the LDC after submitting the following two agreement forms:

1. Agreement Concerning Dissemination of TAC Results *(submit to NIST)*
2. TAC 2010 Evaluation License Agreement *(submit to LDC)*

A link to the user agreement forms and instructions for submitting forms can be found at the KBP 2010 web site (http://nlp.cs.qc.cuny.edu/kbp/2010/registration.html). After submitting both forms, registered RTE-6 teams may contact LDC's Membership Office at ldc@ldc.upenn.edu to request the datasets by catalog number and title.

Title	**LDC Catalog #**	**Date Available**
TAC 2010 RTE-6 KBP Validation Pilot Development Data	LDC2010E32	May 10, 2010
TAC 2010 RTE-6 KBP Validation Pilot Test Data	(TBA)	August 17, 2010

4. DATA SET AND SUBMISSION FORMAT

4.1. Development Set

The following items will be distributed as Development set:

- a directory containing the supporting documents returned by the systems participating in the KBP track (i.e. the T's in the T-H pairs)
- a gold standard, which consists of a single file in the following XML format:

```
<rtekbp_devset>
  <pair id="1" query="SF10" entity_type="per" attribute="age" entailment="NO">
    <entity>Chris Simcox</entity>
    <value>one</value>
    <t>APW_ENG_20051219.0897.LDC2007T07</t>
    <h id="1">Chris Simcox is aged one</h>
    <h id="2">Chris Simcox's age is one</h>
    <h id="3">Chris Simcox is age one</h>
    <h id="4">Chris Simcox is one years old</h>
  </pair>
...
  <pair id="742" query="SF13" entity_type="per" attribute="title"
entailment="YES">
    <entity>Gholam-Ali Haddad-Adel</entity>
    <value>lawmaker</value>
    <t>APW_ENG_20081105.1042.LDC2009T13</t>
    <h id="1">Gholam-Ali Haddad-Adel has the title of lawmaker</h>
    <h id="2">Gholam-Ali Haddad-Adel holds the title of lawmaker</h>
    <h id="3">Gholam-Ali Haddad-Adel is a lawmaker</h>
    <h id="4">Gholam-Ali Haddad-Adel is an lawmaker</h>
  </pair>
...
</rtekbp_devset>
```

In the gold standard:

- each slot filler proposed by KBP systems is represented as a single evaluation item which appears within a single <pair> element
- the element <pair> contains information about the KBP Slot Filling task in the following attributes:
    - `query`, a unique identifier for the slot-filling query for the target entity
    - `entity_type`, the kind of the target entity, e.g. PERSON, ORGANIZATION
    - `attribute`, the target slot to be filled
    - `entailment`, the "YES" or "NO" entailment annotation for the pair, as automatically derived from the KBP assessment value for the slot filler
- the element <entity> contains the name of the target entity
- the element <value> contains the slot filler
- the element <t> (text) contains the link to the source document
- the elements <h> (hypothesis) contain all the H's automatically constructed from the slot filler.

4.2. TEST SET

The Test Set format is the same as the Development Set, except that the `entailment` attribute contained in the <pair> element is left empty as it must be returned by the participating systems. Participants are reminded that the Test Set is blind, and its T-H pairs must not be analyzed before submitting the results.

4.3. SUBMISSION FORMAT

A single xml file in the following format must be submitted for each run, and must list all the pairs proposed in the Test Set, together with their respective entailment judgment:

```
<rtekbp_testset>
  <pair id="1" entailment="NO"/>
  <pair id="2" entailment="NO"/>
...
  <pair id="7325"entailment="YES"/>
...
</rtekbp_testset>
```

Note that each proposed evaluation item (with multiple instantiated H's) must be tagged with a single YES/NO decision. Systems are not required to return an intermediate YES/NO judgment for each instantiated H.

As the set of relations and the types of arguments are known in advance and are the same for the Development and for the Test Set (except for the expansion of location slots into more specific “city”, “state or province” and “country” slots), participants may want to tune their systems for these specific relations. For this reason, two different types of submissions are allowed:

1. one for *generic* RTE systems, for which no manual effort was invested to tailor the generic system to the specific slots (beyond fully automatic training on the Development Set);
2. the second for *manually tailored* systems, where it is allowed to invest additional manual effort to adapt the systems for the specific slots.

Participants are allowed to submit up to 3 runs for each submission type, for a total of 6 runs for each team. Each run name must include both a number identifying the run (1-3) and a string identifying the type of run (“generic”/”tailored”), e.g.:

NAMEOFTEAM_1_generic.xml

NAMEOFTEAM_1_tailored.xml

For both types of submissions, participants are kindly asked to explicitly describe in their system reports all the issues related to targeting the specific slots.

5. 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.

6. SCHEDULE

- May 10: Release of Development Set
- August 17: Release of Test Set
- September 17: Deadline for task submissions
- September 24: Release of individual evaluated results
- October 20: Deadline for systems’ reports

REFERENCES

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.

McNamee, P., Dang, H.T. (2009). Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the TAC Workshop*, Gaithersburg, MD, USA.

Peñas, A.; Rodrigo A.; Sama, V.; Verdejo, F. (2007). Testing the Reasoning for Question Answering Validation. In *Journal of Logic and Computation*. http://logcom.oxfordjournals.org/cgi/reprint/exm072