7TH TEXTUAL ENTAILMENT CHALLENGE @ TAC 2011

KNOWLEDGE BASE POPULATION VALIDATION TASK

Task Guidelines

1. INTRODUCTION

Continuing the effort to bring people from the three TAC communities together and to create a common framework in the field of text understanding, in RTE-6 a new Knowledge Base Population (KBP) Validation Pilot[1] was proposed, based on the TAC KBP Slot Filling Task (McNamee and Dang, 2009). The goal of this task was to show the potential utility of RTE systems for Knowledge Base Population, similar to the goals in the Summarization setting.

Three teams participated in the Pilot, submitting a total of 10 runs. The task proved to be particularly challenging for RTE systems, partly due to the sheer volume of text in the Test Set, which took longer than expected to process, as it contained a high number of unexpectedly long texts which were not present in the Dev Set. Considering that the RTE systems that did manage to handle the unexpectedly large Test Set achieved results that were similar to the performance of systems in the KBP Slot Filling Task, it is worth further investigating this validation task, allowing RTE systems to better tune for the exercise and train to deal with larger amounts of data.

For this reason, the KBP Validation task is proposed again in RTE-7, without significant changes with respect to the RTE-6 Pilot task.

This document provides a definition of the KBP Validation 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 task 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* task 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[2].

[1] For more details see (Bentivogli et al., 2010).

[2] 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).

For example, consider the following slot filler and supporting document returned by a KBP system for the *age* attribute for the target entity "Simon Cowell":

Slot Filler: "47"
Document ID: APW_ENG_20070315.1712.LDC2009T13

If the slot filler is correct, then the document APW_ENG_20070315.1712.LDC2009T13 must entail one or more of the following Hypotheses:

H1: Simon Cowell is aged 47.
H3: Simon Cowell's age is 47.
H4: Simon Cowell is age 47.
H5: Simon Cowell is 47 years old.

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 task, 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 task 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-7 KBP Validation data set is based on the data created for the KBP 2009, 2010 and 2011 *Slot Filling* Task, specifically:

- The RTE-7 Development set consists of over 25,000 T-H pairs from the combined RTE-6 Development and Test sets[3], from which the following pairs have been removed:
  - the pairs generated for the location slots "*place of birth*", "*place of death*", "*residence*", and "*headquarters*", which were present only in the Development set and were replace by more specific slots (e.g., "*city of birth*", "*state or province of birth*", and "*country of birth*"...) in the Test set;
  - the pairs generated for the slot "*other family*", which was present only in the Development set, and was not included in the Test set as it overgenerated 'YES' entailments with respect to KBP "Correct" judgments.
  - the pairs where the T's are web documents.[4]

[3] The data for the RTE-6 Development and Test sets were created from the KBP slot-filling system output and slot-filler assessments from KBP 2009 and 2010 respectively. For more details, see the RTE-6 KBP Validation Pilot guidelines at http://www.nist.gov/tac/2010/RTE/RTE6_KBP_Validation_Pilot_Guidelines.pdf.

- the RTE-7 Test set will be created from corresponding test data from KBP 2011.

The definitions of the slots and guidelines for assessing slot fillers for each of the KBP campaigns are available on the TAC 2011 KBP web site at http://nlp.cs.qc.cuny.edu/kbp/2011/annotation.html .

The creation of the RTE-7 KBP Validation 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[5].

A first manual phase, preliminary to the automatic generation of the H's of the data set, requires that several templates are created for each KBP 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:

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 "*Simon Cowell*" and (iii) a correct slot filler "*British*", we obtain both grammatical and ungrammatical H's within the same evaluation item:

H1: Simon Cowell 's origins are in British.
H2: Simon Cowell comes from British.
H3: Simon Cowell is from British.
H4: Simon Cowell origins are British.
H5: Simon Cowell has British origins.
H6: Simon Cowell is of British origin.

---

[4] The decision to remove the pairs where the T's are web documents was taken based on the fact that Web documents are on average significantly longer and require much more time to process, representing a major issue for RTE-6 systems participating in the KBP Validation task.

[5] 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 task 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.

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[6]. 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, "*Sony BMG"* is considered a correct slot filler for the *employee_of* attribute of the target entity "*Simon Cowell*" if the supporting document contains the text "*Simon Cowell had a contract with Sony BMG until 2009*"; therefore, in the KBP Validation task, the Hypothesis "*Simon Cowell works for Sony BMG*" must be judged to be entailed by the same document.

As the KBP Validation Test Set will be created from the KBP 2011 data, it is important to note that the Test Set and Development Set may differ with respect to the size of the data set and the ratio between positive and negative pairs, depending on the number of KBP 2011 systems' submissions and on their performances.

The RTE-7 KBP Validation data (Development Set and Test Set), are distributed by the Linguistic Data Consortium (LDC). Registered RTE-7 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 2011 RTE Evaluation License Agreement *(submit to LDC)*

A link to the user agreement forms and instructions for registering and submitting forms can be found at the RTE-7 web site (http://www.nist.gov/tac/2011/RTE/registration.html). After submitting both forms, registered RTE-7 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 2011 RTE-7 KBP Validation Task Development Data	LDC2011E29	April 29, 2011
TAC 2011 RTE-7 KBP Validation Task Test Data	LDC2011E30	August 17, 2011

4. DATA SET AND SUBMISSION FORMAT

4.1. DEVELOPMENT SET

The following items will be distributed as Development set:

[6] 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. 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.

- 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="602" 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 type of the target entity, e.g. PERSON, ORGANIZATION
  - `attribute`, the 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, 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. 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

April 29	KBP Validation Task: Release of Development Set
August 17	KBP Validation Task: Release of Test Set
September 16	KBP Validation Task: Deadline for task submissions
September 23	KBP Validation Task: Release of individual evaluated results
October 25	Deadline for system reports (workshop notebook version)
November 14-15:	TAC 2011 workshop in Gaithersburg, Maryland, USA

REFERENCES

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.

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