Slot Filler Validation/Ensembling at TAC 2015
Task Guidelines

Version 1.0, October 7, 2015

Introduction

The goal of Knowledge Base Population (KBP) at TAC 2015 is to promote research in and to evaluate the ability of automated systems to discover information about named entities and incorporate this information into a knowledge base (KB). Cold Start KBP attempts to build an entire knowledge base from scratch, given a knowledge base schema and a collection of source documents from which to extract information to fill the KB. The overall task of Cold Start KB Construction (CSKB) is decomposed into two major tasks: Cold Start Entity Discovery (CSED), where names must be extracted and aligned to entities; and Cold Start Slot Filling (CSSF), which involves extracting attributes (slots) for specific entities.

The Slot Filler Validation (SFV) task at TAC 2015 is a diagnostic task that aims to validating the output of the full Slot Filling systems and eliminates the need for a team to have a full Slot Filling system. For the other tracks in KBP 2015, please visit the KBP 2015 web page: http://www.nist.gov/tac/2015/KBP/.

A variant of the Slot Filler Validation task was originally proposed in the Recognizing Textual Entailment (RTE) track of TAC 2010 and TAC 2011, which aimed to show the potential utility of RTE systems for Knowledge Base Population. In the RTE KBP Validation task, the "text" consisted of the provenance (document) returned with each candidate slot filler, and the "hypothesis" was a natural language expression of the relation between the query entity and the candidate filler; the text entailed the hypothesis if and only if the candidate slot filler was Correct according to the associated provenance. The input to the RTE system was the set of T-H pairs generated from the pooled results of all SF systems, and the evaluation metric was P/R/F1 on the pooled SF output.

In TAC 2013 and 2014, the Slot Filler Validation track was proposed with a slightly different use case and evaluation metric than those in RTE-6 and RTE-7. The 2013 and 2014 SFV track focused on the refinement of output from slot filling systems by either applying more intensive linguistic processing to validate individual candidate slot fillers (as in RTE-6 and RTE-7) *or combining information from multiple slot filling systems*. Evaluation was based on the change in F1 of individual SF systems, rather than on P/R/F1 of the single pool of SF output.

Slot Filler Validation is again proposed at KBP 2015, using output from TAC KBP 2015 Cold Start Slot Filling systems. Two SFV task variants are allowed: the **SFV Filtering task** aims to improve Precision and F1 of each input slot filling run, by removing incorrect candidate slot fillers; the **SFV Ensemble task** aims to produce a single ensemble Cold Start slot filling run by carefully including only selected slot fillers from the input slot filling runs.

This document provides a definition of the KBP 2015 Slot Filler Validation Tasks and a description of the data set, together with instructions on how to take part in the exercise. ***It is assumed that the reader has already read the guidelines for the Cold Start Slot Filling task variant in the***

Cold Start 2015 Task Description
(http://www.nist.gov/tac/2015/KBP/ColdStart/guidelines.html).

Task Description

The slot filler validation (SFV) track is motivated by a use case in which the SFV system is used as a component of a full SF system. It focuses on the refinement of output from Cold Start slot filling (SF) systems by either combining information from multiple slot filling systems, or applying more intensive linguistic processing to validate individual candidate slot fillers.

The Cold Start Slot Filling task of 2015 is focused on searching a collection of newswire and discussion forum documents and extracting values for a pre-defined set of attributes ("slots") for query entities. Given a named entity mention and an attribute for that entity, CSSF systems must find in the document collection the correct value(s) for that attribute and return the slot filler(s) together with its provenance, where provenance is a set of text spans from documents in the corpus that justify the correctness of the slot filler. A CSSF evaluation query is a sequence of one or two slot filling queries to be applied in series (in order to answer questions such as "What schools are attended by the children of Homer Simpson"). A CSSF system must process the evaluation queries in two "rounds", where a Round2 query file is generated from the Round1 output file -- each Round1 slot filler becomes a query entity mention for a Round2 query.

The input to the *Slot Filler Validation* system is a set of submission files from several Cold Start Slot Filling runs (with the run ID anonymized appropriately). The output of the SFV system is a single tab-delimited file with binary classifications (Correct/Incorrect, i.e., 1/-1) of each candidate slot filler in each Cold Start run.

SFV 2015 follows the evaluation procedure of the Cold Start 2015 SF task; namely, if a Round 1 candidate answer from a particular Cold Start run is Wrong, then all the Round 2 candidate fillers *from this Cold Start run* that were generated from the Wrong Round 1 candidate, will automatically be considered Wrong.

For the **SFV Filtering task**, the evaluation measures the effect of using the SFV output to filter the contributing Cold Start runs. Each contributing Cold Start run will be filtered and re-scored in the same way as for the full Cold Start Slot Filling task, and the results compared against the scores for the original unfiltered Cold Start run. SFV tries to increase Precision of the contributing Cold Start runs without significantly reducing Recall, and the objective function for SFV Filtering is to maximize the mean F-score over each of the filtered contributing runs.

For the **SFV Ensemble task**, the evaluation considers the SFV run as a single ensemble Cold Start SF run that includes only selected slot fillers from the input Cold Start SF runs; the objective function for the SFV Ensemble task is to maximize the F-score of the ensemble Cold Start run. In order to compare the F1 score of the SFV ensemble Cold Start run to the F1 score of the input Cold Start SF runs, the SFV ensemble Cold Start run must be careful to exclude redundant slot fillers for the same query (i.e., multiple filler mentions that refer to the same entity or concept).

Input

The Slot Filler Validation data set is based on the runs submitted to the TAC 2015 Cold Start Slot Filling and KB Construction tasks. Each Cold Start run file is in the format of a Cold Start Slot Filling submission file.

1 Input Format

The input to the slot filler validation task consists of 4 types of files:

1) Cold Start SF input
   a. Cold Start Evaluation Source Corpus
   b. Cold Start SF queries (including Round 1 and Round 2)
2) Cold Start output files (in the format of a Cold Start Slot Filling system output file), generated in response to Cold Start SF input
3) Cold Start System profile file (optional)
4) Assessment of all Cold Start system output for a small number of Cold Start Slot Filling queries (optional)

The slot definitions, assessment guidelines, and description of the Cold Start Slot Filling input and output are available under the Cold Start 2015 guidelines (http://www.nist.gov/tac/2015/KBP/ColdStart/guidelines.html).

For the slot-filler validation task, the Cold Start **teams** will be anonymized. The output files of the Cold Start runs will have the run ID in Column 3 replaced by a SFV query ID consisting of the anonymized team_name + task variant (KB or SF) + run_number + filler_candidate_number. Thus, the SFV query ID shows which slot filler candidates came from the same Cold Start run and which Cold Start runs were produced by the same team, for the benefit of SFV teams who may want an idea of system diversity when applying cross-system voting approaches.

2 Training Data

Evaluation data from the KBP 2013-2014 SFV track will be provided to SFV 2015 participants to facilitate system development. The training package will be available on the TAC KBP 2015 website at http://www.nist.gov/tac/2015/KBP/data.html.

Output

For the **SFV Filtering task**, the output of a SFV run should be a single tab-delimited file with the following two fields:

1. SFV query ID of the candidate slot filler
2. Judgment for the candidate slot filler. The possible judgments are:
   1: Candidate slot filler is Correct
   -1: Candidate slot filler is Wrong or Inexact

The definitions of Correct, Wrong, and Inexact for slot fillers are given in the Cold Start 2015 guidelines (http://www.nist.gov/tac/2015/KBP/ColdStart/guidelines.html).

At most one judgment is allowed for each SFV query ID, and a missing judgment for a SFV query ID has a default value of "1" for the SFV Filtering task. An SFV Filtering run is regarded as a filter for each of the input Cold Start SF runs. Applying the SFV filter to a Cold Start SF run may produce an invalid CSSF run, if the filter removed a Round1 candidate slot filler, but neglected to remove the Round2 candidate slot fillers that were generated for this Round1 filler. To ensure that each filtered CSSF run is valid, the organizers will run a script after SFV submission to correct the filtered CSSF run by removing Round2 candidate fillers if their parent Round1 filler had already been removed by the SFV filter.

For the **SFV Ensemble task**, the SFV output should also be a single tab-delimited file with the same two fields as for the SFV Filtering task. At most one judgment is allowed for each SFV query ID, but a missing judgment for a SFV query ID has a default value of "-1" for the SFV Ensemble task. Each SFV Ensemble run is regarded as a single ensemble CSSF run, produced by selecting slot fillers from the set of input Cold Start SF runs. To ensure that each ensemble CSSF run is valid, the organizers will run a script after SFV submission to correct the ensemble CSSF run by removing Round2 candidate fillers if their parent Round1 filler wasn't included by the SFV run in the ensemble CSSF run.

Scoring and Metrics

For the SFV Filtering task, the evaluation metric will be the mean difference in F-measure between the filtered CSSF run and its original unfiltered SF CSSF run. For the SFV Ensemble task, the evaluation metric will be the F1 of the single ensemble CSSF system produced by SFV run.

Submissions

Participants will have at least one week after the evaluation queries are released to return their results. Up to five alternative system runs may be submitted by each team for each task. Submitted runs should be ranked according to their expected score (based on development data, for example). Systems should not be modified once queries are downloaded. Details about submission procedures will be communicated to the track mailing list. The tools to validate formats will be made available at: http://www.nist.gov/tac/2015/KBP/SFValidation/tools.html

Schedule

Please visit the KBP 2015 slot filler validation webpage for the schedule: http://www.nist.gov/tac/2015/KBP/SFValidation/index.html

Mailing List and Website

The Slot Filler Validation track website is http://www.nist.gov/tac/2015/KBP/SFValidation/. The KBP mailing list is tac-kbp@nist.gov. Information about subscribing to the list is available at: http://www.nist.gov/tac/2015/KBP/registration.html.