Task Description for Knowledge-Base Population at TAC 2009

Version of: 2/20/09

Overview

The high-level goal is to have systems to discover information about named entities and incorporate this information in a knowledge source. An initial (or reference) knowledge base will be provided along with a document collection that systems are to use to learn from. Attributes derived from Wikipedia infoboxes will be used to create the reference knowledge base. There will be two related tasks: Entity Linking, where names must be aligned to entities in the KB, and Slot Filling, which involves mining information about entities from text. Groups may participate in either, or both tasks, but participation in both is encouraged.

The tasks will be structured by having participants process a list of target entities. The list will contain entity types such as Person (PER), Organization (ORG), and Geo-Political Entity (GPE); GPEs include inhabited locations such as cities, countries, and continents.

Knowledge Base

Wikipedia infoboxes will be the basis for the reference knowledge base; however exact compliance with Wikipedia is not intended. The KB will be derived from a subset of entities from Wikipedia that have infoboxes, and it will likely be on the order of 100,000 nodes. Each entity in the knowledge base (sometimes called a node) will include the following:

- a name string
- a node ID (a unique identifier)
- a reference to disambiguating text (such as text from the Wikipedia page)
- a set of slot names and values
- an assigned entity type of PER, ORG, or GPE.

For each of the three entity classes a list of desired attributes will be published. Some attributes will be atomic (e.g., *PER:date-of-birth*) while others may be list valued (e.g., *PER:employer*, *PER:children*). The current plan is for the set of slot names, or attributes, for each node to remain unchanged from their representation in Wikipedia. A mapping from original infobox slot-names will be provided. For example, *University:established* and *NFL_Team:founded* should map to *ORG:creation_date*.

Infoboxes are not an ideal knowledge representation; one key disadvantage is a lack of inheritance (and therefore, consistency). Infoboxes also tend to focus on presentation on a Wikipedia page instead of abstract representation. As an example,

consider the table below. The infobox for each of these organizations contains a slot related to the date it was created; however, the name of the slot varies.

Organization	Infobox Class	Slot Name	Slot Value
Bill/Melinda Gates Foundation	Non-Profit	founded_date	1994
Cornell	University	established	1865
FDA	Government_Agency	formed	1906
NASA	Government_Agency	formed	July 29, 1958
Washington Redskins	NFL_Team	founded	1932

Entity Linking

The task is to determine for each target list entry, which entity is being referred to. Target list entries will be of the form [name-string, docid]. Each name-string will occur in the referenced test collection document. The purpose of the associated document is to provide context that might be useful for disambiguating the name-string. Entities will generally occur in the list multiple times using different name variants and/or different docids. It is expected that some entities will share confusable names (e.g., *George Washington* could refer to the president, the university, or the jazz musician; *Washington* could refer to a city, state, or person).

For each list entry a KB node-id must be returned. For entities that have no corresponding node in the initial KB, this should be determined and a new system-generated ID should be created. All name strings in the target list that refer to the same entity should be assigned the same ID. System responses will be evaluated based on the correctness of linkages to KB nodes.

The Slot Filling task (see below) also contains an entity linking component. For example, if in that task a fact is learned such as *birthplace(Paul Newman, Cleveland)* for a Paul Newman entity, then not only should the text "Cleveland" be identified, but that slot value must also be linked to a node for the corresponding GPE, if one exists in the KB.

Results for the entity linking task will be due before the target list for the slot-filling task is provided. This is because the slot filling list will contain correct node-id assignments for target entities. The entity linking task is similar to cross-document co-reference; however, here the problem requires alignment to a knowledge base, not clustering of entities. The B-Cubed metric (Bagga and Baldwin, 1998) will probably be the metric used for evaluation.

Slot Filling

This task involves learning a pre-defined set of relationships and attributes for target entities based on the evaluation corpus. Entries in the slot filling target list will be look like [name-string, docid, entity-type, node-id]. For example: [Paul

Newman, ABC-20080611-9372, PER, KB-2317] might be a target list entry for actor Paul Newman. The node id that is provided will refer to an appropriate node in the KB. For targets for which no node exists in the KB, a node-id like *NIL-001* will be given. As for the entity linking task the provided docid is intended to give context for the entity.

Slots can be of one of two types: those that admit only a single value (e.g., *PER:date-of-birth*) and those can accept more than one value (e.g., *PER:employer*). Systems should fill-in slots using the document collection. For entities previously contained in the KB, redundant information should not be provided; only novel information is of interest. This does not necessarily mean that if an attribute has a value in the initial KB that the slot should be ignored entirely. For example, if the alma-mater slot for investor Warren Buffet looked like:

alma-mater: University of Nebraska

but a document is found which notes that he graduated from Columbia University, then this should be reported. Systems are not expected to correct or modify values from the reference KB, but only to add information. Therefore single-value slots which already have an entry in the KB should not be changed.

System responses will resemble: [node-id, slot-name, answer-string, docid, kb-link]. The node-id will be provided in the input list. The docid should support the answer-string as being correct. The purpose of the kb-link is to provide a node-id for the answering string, if one exists in the KB. If no novel information is learnable for a slot, then a response of [slot-name, NIL] is appropriate. To supply more than one value for multi-valued slots, multiple responses should be provided for the same node-id/slot-name.

The list of slots for each generic class (i.e., Person, Organization, or GPE) will be published along with a description of what appropriate values should look like. Some slots will be applicable to all members of a class (e.g., *PER:date-of-birth*), but others will make sense only for a subset of entities. For example, *ORG:stock-symbol* is only applicable to publicly traded companies.

Correctness of slot answers will be evaluated in a similar fashion as was done in the TREC Question Answering track. In particular, system answers may be marked Correct, Wrong, Inexact, or Redundant (for list-accepting slots). Wrong includes both incorrect and not supported by the cited docid; this is a fusion of Wrong and Unsupported in TREC QA. The metric for scoring slot-filling is still to be determined, but F1-scores averaged across all slots might be a reasonable choice. There may be limitations on the use of human judgments for post hoc experimentation, as was the case with TREC QA. In other words, in a future experiment a correct slot value may be discovered for an entity, but that value might never have been judged in the original evaluation.

Within-KB linkage of slot values will be measured separately. KB-linkage of slot *values* is a co-reference task. However, if a system fails to find a correct answer then it isn't possible to assign a correct link (i.e., there is a cascading errors effect).

Schema

The set of permissible slot values for Persons, Organizations, and Locations is being refined. A notional list is provided below. The final lists will include guidelines for that clarify what values should look like and which slots are atomic vs. multi-valued.

Persons

- Aliases/Nicknames/Variants
- Birth name (if different from name)
- Age
- Birth date (if stated, not computed)
- Birth place
- Home town
- Death date
- Resides-in
- Nationality(ies)
- Schools attended
- Degrees held
- Employer(s)
- Occupation
- Religion
- Spouse
- Parents
- Children
- Siblings
- Email address
- Phone number
- Salary, or Net worth

Organizations

- Aliases/Nicknames/Variants (also previous names)
- Date established
- Location of Headquarters
- Membership size (including employees, or volunteers, or members, as appropriate)
- Leader (could include CEO, 'Director', Coach of a sports team, president, …)
- Stock ticker, if publicly traded company
- Annual revenues, budget, or income
- Motto
- Website

GPEs

- Aliases/Nicknames/Variants (also previous and foreign names)
- Date settled (if appropriate)
- Latitude and Longitude
- State or Province (if applicable)
- Country (if applicable)
- Population (if inhabited region)
- Political leader (if governed)
- Seat of government (i.e., county seat or state/national capital, if appropriate)

Data

More precise details about the format of the target lists, system responses, and data sources will be forthcoming. The document collection will probably be mostly newswire and on the order of a million articles.

Schedule (tentative)

Sample data released: May 1
Reference KB and document collections available: June 10
Entity Linking target list released: July 2
Entity Linking results due: July 9
Slot Filling target list released: July 13
Slot Filling results due: July 20
Assessments available: September 22
TAC workshop: November 16-17

FAQ

1. I have a question about the task.
   Please post it to the list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://apl.jhu.edu/~paulmac/kbp.html
2. Are there any constraints concerning the use of external resources such as internet search, WordNet, and Wikipedia?
   As a general rule, such resources can be used and it is expected that workshop papers include details about how systems make use of them. However, since the evaluation is specifically proposing to use Wikipedia information as an initial knowledge source, teams should avoid using Wikipedia infoboxes to directly fill slots. Additionally, participants are asked to refrain from editing Wikipedia pages for target entities, either during, or after the evaluation. There are many ways in which Wikipedia can be legitimately used, including as training data for attribute learning, or to compile lists of name variation based on hyperlinks and redirects.