An Evaluation of Technologies for Knowledge Base Population

Paul McNamee[1], Hoa Trang Dang[2], Heather Simpson[3], Patrick Schone[4], Stephanie M. Strassel[3]

(1) Johns Hopkins University Human Language Technology Center of Excellence
(2) National Institute of Standards and Technology
(3) Linguistic Data Consortium, University of Pennsylvania
(4) US Department of Defense

paul.mcnamee@jhuapl.edu, hoa.dang@nist.gov, {hsimpson,strassel}@ldc.upenn.edu, pjschon@tycho.ncsc.mil

Abstract

Previous content extraction evaluations have neglected to address problems which complicate the incorporation of extracted information into an existing knowledge base. Previous question answering evaluations have likewise avoided tasks such as explicit disambiguation of target entities and handling a fixed set of questions about entities without previous determination of possible answers. In 2009 NIST conducted a Knowledge Base Population track at its Text Analysis Conference to unite the content extraction and question answering communities and jointly explore some of these issues. This exciting new evaluation attracted 13 teams from 6 countries that submitted results in two tasks, *Entity Linking* and *Slot Filling*. This paper explains the motivation and design of the tasks, describes the language resources that were developed for this evaluation, offers comparisons to previous community evaluations, and briefly summarizes the performance obtained by systems. We also identify relevant issues pertaining to target selection, challenging queries, and performance measures.

1. Introduction

The Text Analysis Conference (TAC) [1] is a NIST-organized community evaluation of Natural Language Processing technologies which began in 2008. At TAC 2009 the Knowledge Base Population (KBP) track was initiated to foster research in automatically extracting information about named-entities from unstructured text and inserting that information into a knowledge base (KB). The TAC-KBP evaluation builds on the work of the Automatic Content Extraction (ACE) (Doddington *et al.*, 2004) and the Question Answering (QA) evaluations of the Text Retrieval Conference (TREC)[2] and TAC.

Over its history, ACE evaluated the extraction of entities, relations, and events in different languages and multiple genres of text and transcribed speech. In recent years, it had also begun to evaluate cross-document co-reference analysis on larger collections of documents (Strassel *et al.,* 2008). Yet ACE never addressed issues involved with directly populating KBs, such as handling millions of documents, KB node disambiguation, and cross-sentential relation discovery.

The TAC-QA evaluation (and its TREC-QA predecessor) addressed temporally-prioritized factoid, list, opinion, and other style questions about focus entities or events whose answers were drawn from large-scale newswire and blog corpora. However, in these QA tracks, entity disambiguation was not a subject of the evaluation, nor was a study of answering a fixed set of questions while varying the focus entities. Lastly, though blogs were recently added to the QA evaluation as a means of identifying information from less formal media, there was less incentive for participants to answer questions from such data rather than from the more formal sources.

The TAC-KBP evaluation was created to address some of these issues. In terms of infrastructure, TAC-KBP was designed to require participants to process a very large collection of documents, and to coordinate that processing by interacting with a large simulated KB. In this first year of the track there were two tasks: Entity Linking, which is concerned with KB node disambiguation – grounding observed surface mentions to a particular entity in the KB; and Slot Filling, which is the detection of previously unknown attributes about entities – which is akin to question answering with a fixed set of questions. We describe each of these aspects of TAC-KBP in detail.

2. Datasets

TAC-KBP made available a new collection of 1.3 million mixed-genre documents and a compilation of information from approximately 818,000 entities covered in Wikipedia. Many Wikipedia pages contain semi-structured *infoboxes*, tables that list attributes about the page's subject. For example, Wikipedia's NASA infobox includes information such as the agency's founding date (July 29, 1958) and the number of its employees (17,900).[3]

The October 2008 snapshot of Wikipedia was processed

[1] http://www.nist.gov/tac/
[2] http://trec.nist.gov/
[3] http://en.wikipedia.org/wiki/NASA. Visited 10/19/09.

to construct a reference KB to support TAC-KBP. For each page containing an infobox, the following information was collected and formatted in XML: page title, infobox class, the set of facts contained in the infobox table, and the article text. The resulting cluster of information for a single extracted page was assigned a unique entity identifier, forming an entity node in the KB.

3. Entity Linking Task

The Entity Linking task requires aligning a textual mention of a named-entity (a person, organization, or geo-political entity)[4] to its appropriate entry in the knowledge base, or correctly determining that the entity does not have an entry in the KB. The problem is complicated by the fact that entities can be referred to using multiple name variants (*e.g.,* aliases, acronyms, misspellings) and that they may share one or more name variants with another distinct entity (*e.g.,* Washington might refer to a person, city, state, or football team).

3.1 Target Selection

3904 queries were developed where each query consisted of a focus named entity mention and a document containing that mention. The associated document provided context with which to disambiguate the entity. The set of target entities was intentionally designed to include many ambiguous names. Additional detail about the selection process can be found in Simpson *et al.* (2010). A breakdown of the queries is given in Table 1, where 'Present' or 'Missing' indicate the entity's status in the KB.

Type	*# Queries*	*Present*	*Missing*
PER	627	255	372
ORG	2710	1013	1697
GPE	567	407	160
All	3904	1675	2229

Table 1: Number of queries by type and KB presence.

57% of queries were absent from the KB; however, as can be seen in Table 1, GPEs are better represented compared to PERs and ORGs due to the fact that Wikipedia has broad coverage of inhabited locations. A few sample queries are shown in Figure 1.

3.2 Evaluation

Mean accuracy across all queries was selected as the official scoring metric for entity linking. Table 2 reports the top and median scores from 35 runs using micro-averages across the 3904 queries, and breaking down performance based on presence in the KB. A baseline of always predicting absence from the KB ("NIL") would have achieved an official score of 0.57, which most systems beat. Table 3 reports macro-averages across distinct entities. Figure 2 shows the range in scores from the top-submitted run from each team.

DeLorean Motor Company (E0784101)
The creation of renowned automotive engineer John DeLorean, **DMC** eventually made fewer than 9,000 cars, ...

Darryl McDaniels (E0079732)
Rapper **DMC** of Run-DMC is 43.

Detroit Medical Center (NIL)
Mike Duggan, Detroit Medical Center President and CEO, said he has scheduled a meeting with the center's medical leadership and the Ford system to learn more. "It was a great initiative on Henry Ford's part," said Duggan, noting he is particularly interested in restricting salespeople's access to **DMC** facilities.

Figure 1: Sample queries for target "DMC". Only a small excerpt from the provided document is shown.

Status	*# Queries*	*Top*	*Median*	*NIL Baseline*
In KB	1675	0.7725	0.6352	0.0000
Missing	2229	0.8919	0.7891	1.0000
All	3904	0.8217	0.7108	0.5710

Table 2: Top and median micro-averaged performance across all submitted runs compared to a NIL baseline.

Status	*# Entities*	*Top*	*Median*	*NIL Baseline*
In KB	182	0.6696	0.5335	0.0000
Missing	378	0.8789	0.7446	1.0000
All	560	0.7704	0.6861	0.6750

Table 3: Top and median macro-averaged performance across all submitted runs compared to a NIL baseline.

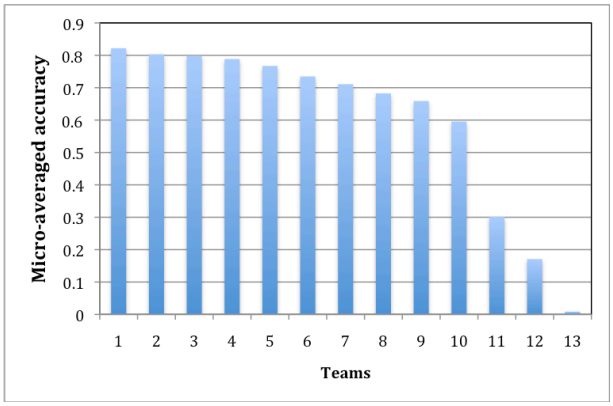

Figure 2: Micro-averaged accuracy of the top-submitted run from each of 13 teams across the 3904 queries.

[4] These are a subset of the ACE taxonomy of entities, and are commonly referred to by PER, ORG, or GPE.

3.3 Hard Queries

Analysis of some of the more difficult queries identified the following phenomena as being particularly troublesome for systems:

- Ambiguous acronyms. Query #1213 "DRC" refers to the Democratic Republic of Congo; however, the provided document uses both DCR and DRC as acronyms.
- Subsidiary organizations. Query #3871 "Xinhua Finance" is referring to Xinhua Finance Media Ltd., but the parent company has a nearly identical name, Xinhua Finance Ltd.
- Metaphorical names. "Iron Lady" is used in query #1717 to refer to Ukrainian prime minister Yulia Tymoshenko.
- Metonyms. The article in query #2599 discusses World Cup rankings and it is difficult to decide whether "New Caledonia" refers to the nation, or its national football team.

3.4 Related Evaluations

A few studies have examined linking named-entities to Wikipedia articles (Bunescu and Pasca, 2006; Cucerzan 2007), but we are unaware of any prior community evaluation in linking entities to knowledge base entries. There have been disambiguation-oriented evaluations that have focused on clustering co-referent entities, including ACE 2008, which contained a cross-document co-reference task, and the Web People Search (WePS) workshops, which have studied clustering Web pages that contain ambiguous person names (Artiles *et al.,* 2008).

4. Slot Filling Task

The Slot Filling task required participants to automatically distill information from the document collection which fills missing KB attributes for focus entities. There were 42 different slots (or attributes): 20 for persons; 14 for organizations, and 8 for geo-political entities.

Different subclasses of entities have type-specific attributes. For example, professional tennis players have attributes such as their height, whether they are right or left-handed, and the amount of money they have won in tournaments. A choice had to be made whether to include narrow features particular to only certain types of entities or whether to focus on generic features that transcend many entity types. It was decided to work with generic slots for the coarse-grained classes of persons, organizations, and geo-political entities.

The task caters equally to the information extraction and question answering communities. Many slots fit naturally with a relation extraction approach, for example *per:title*, *per:spouse*, and *per:employee_of*; however, traditional information extraction systems may find other slots more problematic. A slot such as *org:top_members/employees* imposes additional restrictions beyond a mere employee/employer relationship (*i.e.,* the notion of top-ness) and a slot like *org:shareholders* is a much more specific relation than employment or marriage. Unlike factoid question answering, a QA approach to the slot filling task does not require analyzing the question type because the desired response type for each slot is explicit and known ahead of time.

The attributes to be learned were different for each of the three entity types:

- Persons: alternate names; age; date and place of birth; date, place, and cause of death; national or ethnic origin; places of residence; spouses, children, parents, siblings, and other familial relationships; schools attended; job titles; employers; organizational memberships; religion; criminal charges.
- Organizations: alternate names; political or religious affiliation; top members/employees; number of employees; members; member of; subsidiaries; parents; founder; date founded; date dissolved; location of headquarters; shareholders; website.
- Geo-political entities: alternate names; capital; subsidiary organizations; top employees; active political parties; when established; population; currency.

53 target entities were selected for the task: 17 PERs, 31 ORGs, and 5 GPEs. 20 of the targets had entries in the reference knowledge base.

4.1 Evaluation

Updating or correcting existing facts in the KB was not part of the Slot Filling task. System responses were required to be correct, exact, and not redundant with information already present in the KB profile for the target entity, if one existed. The assessment process and scoring were similar to those used in the TREC QA evaluations of factoid and list questions (Dang *et al.,* 2006).

Each submitted slot value was marked as Correct, Inexact, Redundant, or Wrong. For a slot value to be deemed correct, systems had to provide a single docid that supported the extracted slot value. If the associated docid did not support the slot value, then the response was marked wrong.

Some slots can only contain a single value (*e.g., per:date_of_birth*) while others permit multiple values (*e.g., per:schools_attended*). Accuracy was computed over all single-valued slots. Precision, recall, and an F-score were computed for each list-valued slot, with more weight given to precision. The final SF-value score for the Slot Filling task was computed as the mean of the Accuracy (over single-valued slots) and mean F-score (over list-valued slots).

In the KBP '09 evaluation only 39 of the 255 (15%) single-valued slots were found to have supported values in the corpus. Similarly, of 499 list-valued slots, the judged pools only contain correct values for 129 (26%). This has the somewhat undesirable consequence of letting a baseline approach of always guessing NIL (no response) achieve an accuracy of around 80%.

Prior to the evaluation it was unknown how much extractable information was present in the corpus for each entity. There were discussions about giving less importance to NIL-valued slots; however, when adding information to a KB it is very important not to add erroneous information.

Table 4 reports the top and median scores from the submitted runs, broken down by single or lists slots, and by slots which had a known correct response (*i.e.,* non-NIL slots). Because most slots for most entities were not filled from the corpus, the baseline of not extracting any attributes proved hard to beat.

		Top	*Median*	*NIL*
Single	All 255 slots	0.816	0.514	0.847
	39 non-NIL slots	0.436	0.154	0.000
	215 NIL slots	0.926	0.596	1.000
List	All 499 slots	0.742	0.439	0.741
	129 non-NIL slots	0.292	0.141	0.000
	370 NIL slots	0.926	0.596	1.000
SF-value score		0.779	0.461	0.794

Table 4: Top and median performance for slot filling runs compared to a 'no prediction' baseline.

5. Summary

The KBP 2009 evaluation was motivated by a desire to improve extraction and question answering technologies in the context of adding information to an existing KB. The entity linking task studied grounding name mentions to specific knowledge base entries. The top system achieved an accuracy of 82% on the dataset. The slot filling task was a departure from the document-centric model used at previous evaluations such as ACE. There are challenges that still need to be addressed in evaluating slot filling, including: coping with the sparseness of learnable, novel facts produced by systems, and working with generic entity categories.

Future directions could include temporally qualifying assertions, detecting changes in KB values or contradictions, and extracting information from languages other than English. In the meantime these test collections will serve as useful benchmarks for these tasks.

The resources described within this paper will be made available to the larger research community after the conclusion of the KBP 2009 evaluation. Source data, annotations, scoring software and related linguistic resources will be published in the LDC catalog as an integrated KBP 2009 evaluation corpus. Other resources including KBP system descriptions, a track overview paper (McNamee and Dang, 2010), and site papers will be published on the NIST TAC website.

References

Artiles, J., Sekine, S., and J. Gonzalo (2008). Web People Search: results of the first evaluation and the plan for the second. Proceedings of the 17th International Conference on the World Wide Web, pp. 1071-1072.

Bunescu, R. C. and M. Pasca (2006). Using encyclopedic knowledge for named-entity disambiguation. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.

Cucerzan, S. (2007). Large-scale named-entity disambiguation based on Wikipedia data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Dang, H. T., Lin, J., and D. Kelly (2006). Overview of the TREC 2006 Question Answering Track. Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006).

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and R. Weischedel (2004). Automatic Content Extraction (ACE) Program: task definitions and performance measures. Proceedings of the Fourth International Language Resources and Evaluation Conference (LREC '04).

McNamee, P. and H. T. Dang (2010), Overview of the TAC 2009 Knowledge Base Population Track. To appear in the Proceedings of the Second Text Analysis Conference (TAC 2009).

Simpson, H., Parker, R., Strassel, S., Dang, H. T., and P. McNamee (2010). Wikipedia and the Web of Confusable Entities: experience from Entity Profile creation in TAC Knowledge Base Population. Proceedings of the Seventh International Language Resources and Evaluation Conference (LREC '10).

Strassel, S., Przybocki, M., Peterson, K., Song, Z., and K. Maeda (2008). Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC '08).