Wikipedia and the Web of Confusable Entities: Experience from Entity Linking Query Creation for TAC 2009 Knowledge Base Population

Heather Simpson[1], Stephanie Strassel[1], Robert Parker[1], Paul McNamee[2]

(1) Linguistic Data Consortium, University of Pennsylvania
3600 Market St., Suite 810, Philadelphia PA, 19104, U.S.A.
(2) Johns Hopkins University Human Language Technology Center of Excellence
E-mail: {hsimpson, strassel, parkerrl}@ldc.upenn.edu, paul.mcnamee@jhuapl.edu

Abstract

The Text Analysis Conference (TAC) is a series of Natural Language Processing evaluation workshops organized by the National Institute of Standards and Technology. The Knowledge Base Population (KBP) track at TAC 2009, a hybrid descendant of the TREC Question Answering track and the Automated Content Extraction (ACE) evaluation program, is designed to support development of systems that are capable of automatically populating a knowledge base with information about entities mined from unstructured text. An important component of the KBP evaluation is the Entity Linking task, where systems must accurately associate text mentions of unknown Person (PER), Organization (ORG), and Geopolitical (GPE) names to entries in a knowledge base. Linguistic Data Consortium (LDC) at the University of Pennsylvania creates and distributes linguistic resources including data, annotations, system assessment, tools and specifications for the TAC KBP evaluations. This paper describes the 2009 resource creation efforts, with particular focus on the selection and development of named entity mentions for the Entity Linking task evaluation.

1. Introduction

The Text Analysis Conference (TAC) is a series of Natural Language Processing evaluation workshops organized by the National Institute of Standards and Technology (NIST). In 2009, TAC added a Knowledge Base Population (KBP) Track, designed to support development of systems that are capable of automatically populating a knowledge base with information about named entities mined from unstructured text. (McNamee et al. 2010). KBP is a hybrid descendant of two evaluation programs: TREC Question Answering (Dang et al. 2006) and Automated Content Extraction (ACE) (Doddington et al. 2004). TAC 2009 KBP evaluated systems on two main tasks, the Entity Linking task and the Slot Filling task. The Entity Linking task required systems to accurately associate text mentions of unknown person (PER), organization (ORG), and geopolitical (GPE) names to entries in an external knowledge base. This task is somewhat similar to the WePS2 Clustering task (Second Web People Search Evaluation Workshop 2009), which provides systems with a set of person names and expects them to cluster web search results for those names by their reference to a unique person entity. The Slot Filling task required systems to populate Wikipedia-style infoboxes for a set of specific entities with information found in the 2009 source data.

Linguistic Data Consortium (LDC) at the University of Pennsylvania creates and distributes linguistic resources including data, annotations, system assessment, tools and specifications for TAC KBP. This paper describes the process of resource creation for the TAC 2009 KBP Entity Linking evaluation. We first describe the identification of Person, Organization and Geopolitical entities of interest based on a set of "seed" entities. We then discuss the realization of those entities in the TAC KBP test corpus, and provide information about the treatment of those entities in Wikipedia, which serves as the knowledge base for TAC KBP.

2. Resource Creation for Entity Linking

The 2009 KBP Slot Filling and Entity Linking evaluations used a source data corpus and knowledge base provided by LDC. The source data was selected from LDC's existing English-language collections of newswire articles, as well as a small amount of web data and audio transcripts. The knowledge base was created by compiling information from an October 2008 snapshot of Wikipedia by deriving and formatting the page title, infobox class, infobox information, and article text from more than 800,000 entries, each ostensibly corresponding to a unique entity. Wikipedia infoboxes are semi-structured tables listing facts about that entry's entity. KBP knowledge base entries were created only for Wikipedia entries containing an infobox. A small percentage of the Wikipedia snapshot infoboxes had formatting abnormalities that made their infoboxes difficult to parse, these were also left out of the knowledge base.

For the TAC KBP 2009 Entity Linking task, LDC created the evaluation queries and gold standard answers. Entity Linking queries consist of two elements: a text string corresponding to a name mention of a PER, ORG, or GPE entity, and the id of a document from the KBP source data containing that mention. For each query, systems participating in the Entity Linking task are required to provide a correct link to the named entity's entry in the knowledge base, or correctly report that it does not have an entry in the knowledge base. LDC created a gold-standard mapping for each query to a unique entity id corresponding either to an entry in the knowledge base, or, if the named entity was not represented in the knowledge base, to an entry for that

entity in an internal database.

2.1 Approach

LDC's development of the TAC KBP Entity Linking queries and gold standard mapping required two foundational deliverables:

1. a carefully selected set of PER, ORG, and GPE entities, with some variety in mention frequency, and a large proportion having some interesting name features, such as multiple spellings, and, of particular interest, sharing a name with another entity (confusability)
2. a large, varied text corpus containing sufficient mentions of those entities

These two deliverables and their desirable qualities are nearly identical to those created by LDC in support of evaluation entity and document selection for the ACE 2008 XDOC task.

In ACE 2008, the program goal was to extract information for all entities and relations of targeted types, and perform cross-document co-reference for PER and ORG entities. The entity and corpora selection goals were accomplished using a two-stage process. In the first stage, a large set of candidate entities were selected for entity profile creation by annotators using world knowledge and a selection of existing LDC English text corpora. In the second stage, annotators manually queried a large data pool for the selected entities, and logged information about their representation in the data. Based on the collected information, a set of 250 entities with the desired properties was selected, and a corpus of approximately 10,000 documents was carefully selected to provide rich coverage of those 250 entities. From that set of entities and source document, 50 entities, and 400 documents maximizing coverage of those entities were chosen for use in the ACE 2008 evaluation (Strassel et al. 2008).

In TAC 2009 KBP, the ACE 2008 10,000 corpus was taken as the base for the evaluation source data, but to reach the goal of testing system performance on a large-scale corpus, the corpus was augmented significantly with newswire from LDC existing collection. The TAC 2009 KBP source data corpus contains 1.3 million documents in total. The documents selected for ACE 2008 spanned an epoch from 05/1994 to 12/2006. The epoch of the additional newswire documents, from 01/2007 – 12/2008, was chosen to be concurrent with the 10/2008 epoch of the knowledge base. LDC took advantage of the significantly larger set of source data to avoid the time-consuming process of selecting the corpus for maximum coverage of specific entities. Instead, it was conjectured that a corpus of that size would contain enough density and variety in entity coverage that candidate evaluation entity selection could be done without reference to the corpus.

For selection of the Entity Linking evaluation queries, LDC followed a two-stage process similar to that used in ACE 2008 to select candidate named entities. In the first stage, a large set of candidate entities were selected, and in the second stage matched to their representation in the source data. However, there were a couple major differences in the TAC 2009 KBP approach.

First, the TAC KBP effort focused the process of entity selection by beginning with a small set of "seed" entities containing strong representation of desired qualities for KBP evaluation entities, and expanding on those by adding entities from in the October 2008 Wikipedia snapshot from which the knowledge base was derived. Entities with a Wikipedia entry tend to be newsworthy entities; in fact, entries are required to comply with the Wikipedia notability guidelines (English Wikipedia, 2010). Choosing candidate evaluation entities from Wikipedia, therefore, increases the chance that they will be represented in a large corpus composed mainly of newswire articles such as the 2009 KBP source data.

Second, entity profiles, concise annotator-generated pieces of information meant to uniquely identify an entity, were key to TAC KBP resource creation. Entity profiles were developed for ACE 2008, and used mainly to log name variants used for corpus selection. In TAC KBP, profiles assumed a more central role. In addition to logging name variants to match in the source data, entity profiles served as a connecting thread to uniquely identify entities in initial candidate selection, expansion through Wikipedia exploration, and corpus exploration.

2.2 Seed Entity Selection

The 56 PER and ORG seed entities used in the first stage of Entity Linking query selection were selected from the entities used to select the ACE 2008 source data. Since the ACE 2008 entities had been carefully chosen for properties desirable for KBP such as confusability, they provided a solid foundation for the Entity Linking candidate set. The 16 GPE entities were selected from the Wikipedia entries for the PER and ORG entities. Including the ACE 2008 10,000 documents selected for coverage of those entities in the KBP source data guaranteed that some of the candidate entities would have representation in the source data.

In addition to the shared ACE/KBP desirable characteristics of confusability, name variance, and variety of frequency, another targeted quality of the KBP Entity linking query set was to have some entities with known correct links in the external knowledge base, and some entities with no representation in the external knowledge base (KB). Entities could have three levels of representation with respect to the October 2008 Wikipedia snapshot:

1. Entry in the knowledge base (Wikipedia

snapshot entry containing infobox)

2. Entry in Wikipedia but not in the knowledge base (Wikipedia snapshot entry did not contain an infobox, or in a small set of cases, contained an improperly formatted infobox that could not be parsed).
3. Entry not in Wikipedia or in knowledge base

The seed entities were selected in part for variety in Wikipedia snapshot representation, ensuring some variety in KB representation in the final set of entities. The breakdown by entity type and KB representation of the 72 seed entities is represented in the table below:

	PER	ORG	GPE	Total by KB representation
KB	11	11	15	37
Wikipedia, no KB	10	10	1	21
No Wikipedia	6	8	0	14
Total by entity type	27	29	16	72

Table 1: Seed Entities by Wikipedia snapshot representation and Entity Type

2.3 Entity Profiles

Supporting TAC KBP required LDC to create links between entities with imperfect overlap in representation between the Wikipedia snapshot, the knowledge base, and the corpus, so it was necessary to reference entities independently of these resources. Entity profiles served this purpose in TAC KBP, acting as a portable reference to a unique entity id that was provided in LDC annotation tools as a reference for linking entities with corpus mentions and the knowledge base.

Entity profiles contain one canonical name variant for that entity used as the title, or 'handle' of the profile, entity type classification, possible or likely name variants for the entity, and facts about the entity, for example:

Name: Lincoln County, Arkansas
Entity Type: GPE
Name Variants: Lincoln County, Lincoln
Facts: Lincoln County is a county located in the U.S. state of Arkansas and is included in the Pine Bluff Metropolitan Statistical Area. As of 2000, the population is 14,492. The county seat is Star City.

The Facts field is meant to contain information that would disambiguate that entity from other confusable entities. Annotators were instructed to take that information from the entity's Wikipedia entry if it had one, and if not, from corpus documents. Additional supplementation from external online searching was also permitted.

2.4 Seed Entity Expansion

The goal for the final set of Entity Linking queries was several thousand queries, corresponding to several thousand name mentions of PER, ORG, and GPE entities, with high levels of confusability and variety. Based on experience in ACE 2008, LDC judged that confusability was the most difficult to find of the desirable entity set characteristics. LDC designed the seed entity expansion approach to specifically target confusable entities. Annotators expanded the seed entities organically based on rules of confusability resulting in clusters or webs of confusable entities.

Within the target entity types, annotators were instructed to avoid fictional entities (e.g. "Batman"), non-individual PER entities (e.g. "Hmong"), and 'time-sensitive' entities (e.g. "the 2008 Boston Red Sox", "the 2002 Russian Gymnastics team").

2.4.1 Wikipedia Exploration

Seed entity expansion was accomplished mainly in the first stage of TAC KBP, referred to as Wikipedia Exploration. The task for annotators was to search the Wikipedia snapshot for their assigned entity, add any new name variants or facts to the entity profile, match it to a Wikipedia entry if appropriate, and create new entity profiles for any "confusable" entities they found. LDC developed a tool customized for the Wikipedia Exploration task to provide all the required functionality.

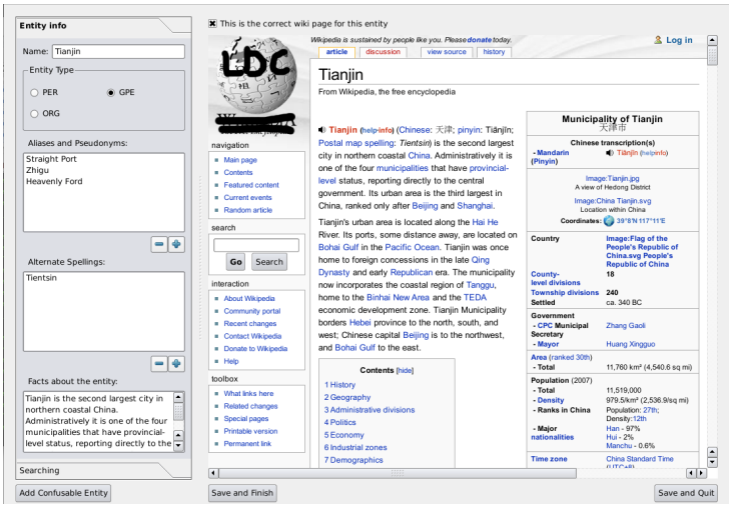

Figure 1: Wikipedia Exploration tool

Annotators were instructed that entities are confusable if they are known or likely to be referred to by the same name variant. For example, "Chicago White Sox", "(the city of) Chicago", and "University of Chicago" are confusable, because they may all be referred to as simply "Chicago". "Chicago" the movie would also be confusable, but not one of the targeted entity types, so it would not be added.

New confusable entity profiles were associated to a "confusable cluster" for the original entity profile. For a single cluster, annotators were instructed to add confusables up to roughly a 2nd or 3rd degree relation from the original entity. There was no hard limit set, to allow annotators to continue with particularly productive clusters. This process turned the clusters into a series of small confusable entity webs based off of each seed entity, where all entities shared at least one name variant with at least one other profile, but there could be varying degrees of overlap.

For example, to the "Chicago" confusable cluster for assigned entity "Chicago (the city)", the annotator could add "University of Chicago", and then "University of California", because it shares the variant "UC" or "U of C". They may then add all the organizations that have the acronym "UC". Most of the "UC" entities would have a 2nd degree relation to "Chicago" because they do not share a variant with "Chicago (the city)", but are connected to it through "University of Chicago", which has the "Chicago" and "UC" variants.

2.4.2 Corpus Exploration

The Wikipedia Exploration task resulted in identification of over 3,000 name variants. LDC performed exact string match on these name variants in the 2009 KBP source data. More than 1,000 variants were found to have document matches in the corpus. These documents, their matching variants, and entity profiles containing those variants were assigned to annotators in the Corpus Exploration tool. For each variant, annotators matched up to name variant/document pairs to the entity profile referred to by the name mention in that document context.

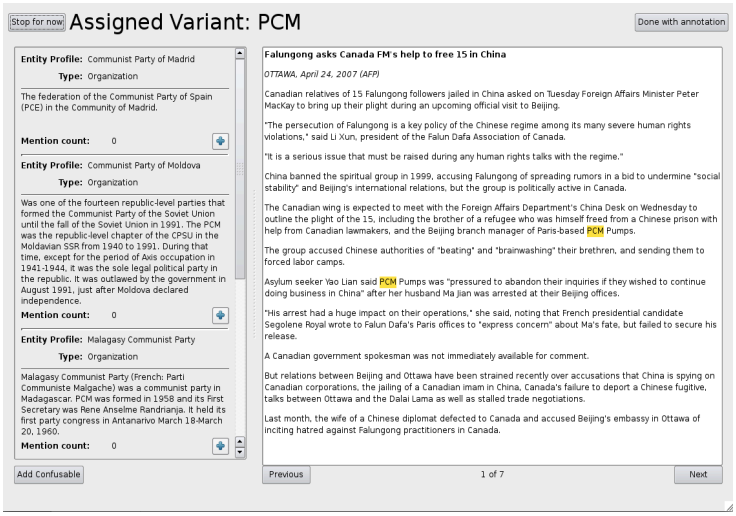

Figure 2: Corpus Exploration tool screenshot

If a name mention occurred which could not be matched to an existing profile, a new entity profile was created for that entity.

3. Resulting Coverage in Wikipedia and Source Data

From the 72 seed entities provided as input to the Wikipedia Exploration task, expansion based on confusability resulted in about 450 unique entities, a 625% increase.

After the Wikipedia exploration and Corpus exploration tasks were completed, the final set of TAC KBP Entity Linking evaluation queries is a set of 3904 document-id and name variant pairs, where a name variant corresponds to an named mention of a Person, Organization, or GPE entity with at least one string match in the corresponding document. There are 560 unique entities in the final Entity Linking target list.

Entity type distribution was not controlled in the selection of the Entity Linking queries, instead we

included all appropriate confusable entities and associated name variants that had matches in the corpus. Thus, the distribution of entity type in the final set of queries in comparison with the original set of seed entities may be used to give some evidence of the coverage in newswire data of confusable entities in Wikipedia by entity type. The distribution of entity type in the original set of 72 seed entities was roughly 40% Person, 40% Organization, 20% GPE. The entity type distribution for entities in the resulting Target Entity linking list is roughly 15% Person, 70% Organization, 15% GPE.

The disproportionate gain in organization entities via the confusable cluster expansion approach may be attributed to a combination of the highly productive nature of organization confusable clusters, as well as potentially richer coverage of confusable organizations in newswire data. Organization confusable clusters may be particularly productive due to the high percentage of organization entities referred to by an acronym, or by metonymy with a location name, and therefore the number of unique ORG entities sharing a name variant will be proportionally higher. A smaller percentage of GPE entities may share a name with multiple other GPEs, but the similarly named GPEs are often unlikely to occur in a corpus of newswire data, since they will only be mentioned if they have participated in a newsworthy event (thus, "Paris, France" will have a lot of information in the corpus, not so for "Paris, Texas"). Person entities sharing more than just a first or just a last name are less common for the same reason – though there may be many people with the same name in the world, only a very small percentage may be newsworthy. A potential solution to balance out confusable entity coverage would be to add more coverage of non-news sources, such as weblog data, to the source data corpus. Weblog data would provide coverage of a greater variety of topics, and also likely include mentions of non-newsworthy entities that are relevant to the blogger's personal life.

Of the 560 Entity Linking target entities, selected for confusability and high 32.5% had a 10/2008 Wikipedia entry with infobox, 33.4% had a 10/2008 Wikipedia entry with no infobox or an unparseable infobox, and 34.1% did not have a 10/2008 Wikipedia entry. This may provide some evidence for extrapolating overall coverage of PER, ORG, and GPE entities in Wikipedia in concurrent newswire data.

4. Conclusion

LDC provided TAC KBP Entity Linking participants with a set of evaluation target entities varied in type, confusability, and representation in the external knowledge base. The process of creating gold standard links to unique entity ids allowed evaluation of system performance with no further human intervention, while simultaneously providing data on the representation of the target entities in the corpus. This corpus supports testing and evaluation of the task of linking named entity mentions to a Wikipedia-derived knowledge base, and could provide useful data on the overlap in representation of newsworthy person, organization, and geopolitical entities in Wikipedia. Further analysis of interest could be to compare the entity coverage in the October 2008 snapshot of Wikipedia with coverage in a newer archived version, adding more mentions taken from non-newswire genres, and investigating inter-annotator agreement. 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 and site papers will be published on the NIST TAC website.

5. References

Dang, H.T., Lin J., Kelly, D. (2006). Overview of the TREC 2006 Question Answering Track. In *Proceedings of TREC 2006*.

Doddington, G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel, R. (2004). Automatic Content Extraction (ACE) program - task definitions and performance measures. In *Proceedings of the Fourth International Language Resources and Evaluation Conference*.

English Wikipedia. (2010). Notability Guidelines. 17 March 2010 http://en.wikipedia.org/wiki/Wikipedia:Notability

McNamee, P., Dang, H.T., Simpson, H., Schone, P., Strassel, S. (2010). An Evaluation of Technologies for Knowledge Base Population. In *Proceedings of the Seventh International Language Resources and Evaluation Conference.*

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

National Institute of Standards and Technologies. (2009). TAC 2009 Workshop. 08 Nov 2009 <http://www.nist.gov/tac/>

Second Web People Search Evaluation Workshop. (2009). Web People Search Task (WePS). 08 Nov 2009. <http://nlp.uned.es/weps/weps-2-task-description/>