Entity Recognition and Disambiguation Challenge Workshop

I. Motivation

With the emerging focus of search engines on semantic search, there is a growing need to understand queries and documents not only syntactically, but semantically as well. Over the recent years, major search engines have redesigned their search results to accommodate some semantic information, particularly recognized entities in queries and search results. Recent information retrieval studies in SIGIR have also paid significant amount of attention on entity-related research. However, techniques for accurate entity recognition and disambiguation are still far from perfect. The motivation of the workshop is to advance the state of the art on entity recognition and disambiguation for both long and short web documents.

The objective of an Entity Recognition and Disambiguation (ERD) system is to recognize mentions of entities in a given text, disambiguate them, and map them to the entities in a given entity collection or knowledge base. Building a good ERD system is challenging because

- Entities may appear as different surface forms.
- The context in which a surface form appears often constrains valid entity interpretations.
- An ambiguous surface form may match multiple entity interpretations, especially in short text.

Importantly, the workshop will be driven by a public ERD challenge. The Challenge is composed of two parallel tracks. In the “long text” track, the challenge targets are pages crawled from the Web; these contain documents that are meant to be easily understandable by humans. The “short text” track, on the other hand, consists of web search queries that are intended for a machine. As a result, the text is typically short and often lacks proper punctuation and capitalization.

Based on the public challenge, we believe that this workshop can provide researchers interested in the field with an opportunity to compare different approaches, exchange thoughts, discuss other perspectives, and formulate a shared vision for better entity research in the future.

II. Challenge Task Definition

We list the high level task definition here and cover additional details about the challenge in the appendix. Given an input document and a knowledge base, the goal of an ERD system is to first recognize the mentions and then link (disambiguate) them to the entities in the knowledge base according to the context. Detailed definitions for these two tracks follow below. We will only focus on the English language for the ERD competition this year.
Short text track
In the short text track, given a short fragment of text, we ask ERD systems to provide a set of valid entity linking interpretations. For each query, the ERD system should use all available context to produce a set of valid entity linking interpretations.

a. Query: total recall arnold schwarzenegger
The phrase “total recall” should be linked to the 1990 movie and “arnold schwarzenegger” should be linked to the actor/governor. Therefore, there is only one valid interpretation that contains two entities here.

b. Query: total recall movie
The phrase “total recall” can be linked to either the 2012 or 1990 movie. The phrase “movie” is not linked as we have excluded the entity Film from the database. Thus, there are two valid interpretations here. Each interpretation contains only one entity.

The ERD systems need to also handle aliases. For example:

c. Query: the governator
The phrase “the governator” should be linked to a TV program by the name as well as Arnold Schwarzenegger, referring to the actor (note that “the governator” is a famous nickname for Arnold Schwarzenegger). There are two valid interpretations here as well.

Long text track
In the long text track competitors face a similar problem, except the input text is longer (e.g., an entire web page). The task again is to identify all the mentions of entities from our reference database that are consistent with the context. For example, given the text "The Governator, as Schwarzenegger came to be known, helped bring about the state's primary election system", the phrase "The Governator" should be linked to the person but not the TV program.

III. Differences from TAC KBP Entity Linking and other workshops
To our knowledge, there exist two workshops that host related challenges. However, we believe that our challenge differs in that we evaluate against settings that are as close to real world scenarios as possible, in particular on web documents and search queries.

Among the workshops, the most popular one is on the TAC KBP shared task. There are several important differences between the ERD challenge and the entity linking track of TAC KBP shared task:

- **Knowledge base and entity types**
The KBP shared task uses an October 2008 dump of English Wikipedia database and only focuses on entities of type Person, Organization, and Location. The ERD challenge targets a more recent snapshot of the Freebase knowledge base covering a wider set of entity types. As Freebase comes with entity relations, it would be interesting to see how participants make use of the structured information in an ERD system.
• **Evaluation**
In KBP, an evaluation entry consists of a string mention (and its offset) within a document. For example, consider the following snippets (from TAC 2011), where the mention is “Alexandria”: The task is for the system to come up with the entity that the word “Alexandria” links to.

Thus it was through Saint John Damascene that the advanced sciences made their apparition among the Arab Moslems, who had burnt the library of Alexandria in Egypt; it was not the Moslems who instructed the Christians, as was believed for some time in Europe.

In ERD, the challenge evaluates the end-to-end performance of the system. No string mention will be provided. Participating systems need to come up with the mentions as well as the entities. It is our intention to evaluate the ability to find mentions of more diverse entity types, such as Books and TV Shows. Please see the detailed evaluation formulation in the [appendix](#).

• **Short text and web documents**
In ERD, we include the task of analyzing short text. Techniques that work well for long documents may not generalize to short documents. We aim to encourage the community to focus more on informal texts such as search queries. As such texts are often ambiguous, we also evaluate multiple hypotheses.

For long documents, our focus will be primarily on the web documents, which covers great variety of writing styles and different signals. Hence, traditional NLP techniques might need to be altered to handle web documents robustly.

Another related workshop is the Microblog workshop¹, which contains a shared task that with focus on the microblogs only. In contrast, the ERD challenge focus on web documents and search queries.

The TREC Knowledge Base Acceleration workshop is also worth mentioning. In the workshop, the CCR: Cumulative Citation Recommendation task is the task of finding relevant articles that are worth citing for a fixed collection of entities. The workshop is related to our challenge. However, the CCR task does not evaluate on the entity linking performance directly.

IV. **Theme and purpose of the workshop.**
The main goal of the workshop is not to create yet another separate community but to create collaborations between existing communities. We welcome and hope participants in TREC and TAC shared task to participate in our challenge.

We also hope the workshop will bring together researchers and practitioners from both industry and academia with variety of background. From the industrial perspective, we prepare the challenge using data close to real world settings. From the academic perspective, we are interested in how to improve

---

the state of the art by blending techniques from the fields of Information Retrieval, Machine Learning and Natural Language Processing.

Given this workshop is organized as a challenge, the main theme of the workshop would be on the techniques for building state-of-the-art ERD systems. All participants of the challenge will be asked to submit output of their actual systems, and will be expected to submit reports and present their systems at the workshop. Given that entity linking and disambiguation is a cross-discipline task that covers a wide range of topics, we expect participants to discuss variety of IR, NLP and ML techniques.

V. Planned Activities

We expect the workshop to take one full day, starting at (approximately) 9:00 am and ending at (approximately) 5:00 pm. We plan to host two invited talks: one in the morning and one in the afternoon. Possible themes for the keynotes include: special considerations in applying entity related techniques on large scale web documents, entity linking applications, and bringing structured and unstructured data together to help ERD systems. Regarding the challenge, a summary of the challenge results will be reported first, followed by talks from participant in each track. If there are too many participants, we will hold a one to two hour poster session to enable all participants an opportunity to present.

VI. Selection process for participants

We will first publish a call for participation in a number of IR and NLP distribution lists. The challenge will attract people from different areas, as each participant will be asked to provide an oral or poster presentation in the workshop. A program committee from researchers with a track record of publishing and/or conducting research in research areas related to entities will be formed. Program committee members will review the submitted papers and provide comments. The reviewers will also choose the papers that could initiate future discussions for oral presentation. If a large number of high quality papers are submitted we will consider allocating time for a poster session where some of the papers will be presented in a different format.

VII. List of organizers (ordered by last name)

David Carmel, Yahoo! Research
Ming-Wei Chang, Microsoft Research
Evgeniy Gabrilovich, Google
Bo-June (Paul) Hsu, Microsoft Research
Kuansan Wang, Microsoft Research

VIII. Names of potential program committee members

Dan Roth, University of Illinois; Heng Ji, Rensselaer Polytechnic Institute; Silviu-Petru Cucerzan, Microsoft Research; Edgar Meij, Yahoo Barcelona
Appendix: Entity Recognition and Disambiguation Challenge

I. Task Definition

Given an input document and a knowledge base, the goal of an ERD system is to first recognize the mentions and then link (disambiguate) them to the entities in the knowledge base according to the context.

In order to study the role of structural information for this task, we choose to use both Freebase and Wikipedia as our knowledge base. As there are many general concept entities in Freebase (e.g., “triangle” or “democracy”), we use a subset of Freebase that places more focus on “proper noun entities” rather than general concepts for evaluation purposes. We called this subset the reference database. For example, “snow” as a form of ice is not in our reference database, but films with the title “snow” are.

Detailed definitions for these two tracks follow below. We will only focus on English language this time.

Short text track
In the short text track, given a short fragment of text, we ask ERD systems to provide a set of valid entity linking interpretations. For each query, the ERD system should use the all available context to produce a set of valid entity linking interpretations.

For example, consider an entity database with the following entries:

<table>
<thead>
<tr>
<th>Freebase ID</th>
<th>Name</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>/m/0gvrws1</td>
<td>“Total Recall”@en</td>
<td>2012 film</td>
</tr>
<tr>
<td>/m/05xftl</td>
<td>“Total Recall 2070”@en</td>
<td>TV show</td>
</tr>
<tr>
<td>/m/0n3zb6d</td>
<td>“Total Recall”@en</td>
<td>TV show</td>
</tr>
<tr>
<td>/m/0fd4x</td>
<td>“Total Recall”@en</td>
<td>1990 film</td>
</tr>
<tr>
<td>/m/0tc7</td>
<td>“Arnold Schwarzenegger”@en</td>
<td>Actor, Politician</td>
</tr>
<tr>
<td>/m/0gj5qr6</td>
<td>“The Governator”@en</td>
<td>TV show</td>
</tr>
</tbody>
</table>

d. Query: total recall arnold schwarzenegger
The phrase “total recall” should be linked to the 1990 movie and “arnold schwarzenegger” should be linked to the actor/governor. Therefore, there is only one valid interpretation that contains two entities here. More precisely, the system should return all valid interpretations in the query in the following format:

QueryID InterpretationSet PrimaryID MentionText Score (tab delimited)
QueryID represents the original query id. InterpretationSet represents the index of the interpretation. Given that there is only one interpretation for the whole query here, the interpretation ID will be always 0 here. MentionText is the surface form that is linked to the entity, and the PrimaryID is the entity ID. The Score field is a score that represents how confident your system is on this entity assignment. The Score field will only be used for analysis purpose, and will not be used in the evaluation.)

For example, for the query “total recall arnold schwarzenegger” (assume the query id is Q01), the system should return:

<table>
<thead>
<tr>
<th>DocID</th>
<th>BeginOffset</th>
<th>EndOffset</th>
<th>PrimaryID</th>
<th>MentionText</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q01</td>
<td>0</td>
<td>/m/0fd4x</td>
<td>total recall</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Q01</td>
<td>0</td>
<td>/m/0tc7</td>
<td>arnold schwarzenegger</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

e. Query: total recall movie
The phrase “total recall” can be linked to either the 2012 or 1990 movie. The phrase “movie” is not linked as we have excluded the entity Film from the database. Thus, there are two valid interpretations here. Each interpretation contains only one entity. Assume the query id is Q02:

<table>
<thead>
<tr>
<th>DocID</th>
<th>BeginOffset</th>
<th>EndOffset</th>
<th>PrimaryID</th>
<th>MentionText</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q02</td>
<td>0</td>
<td>/m/0fd4x</td>
<td>total recall</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Q02</td>
<td>1</td>
<td>/m/0gvrws1</td>
<td>total recall</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

The ERD systems need to also handle alias. For example:

f. Query: the governator
The phrase “the governator” should be linked to a TV program by the name as well as Arnold Schwarzenegger, referring to the actor (note that “the governator” is a famous nickname for Arnold Schwarzenegger). There are two valid interpretations here as well. Assume the query id is Q03:

<table>
<thead>
<tr>
<th>DocID</th>
<th>BeginOffset</th>
<th>EndOffset</th>
<th>PrimaryID</th>
<th>MentionText</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q03</td>
<td>0</td>
<td>the governor /m/0tc7</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q03</td>
<td>1</td>
<td>the governor /m/0gj5qr6</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Long text track
In the long text track participants face a similar problem, except the input text is longer (e.g., an entire web page). The task again is to identify all the mentions of entities from our reference database that are consistent with the context. For example, given the text "The Governor, as Schwarzenegger came to be known, helped bring about the state's primary election system", the phrase "The Governor" should be linked to the person but not the TV program. Assume the document ID The system should return linked mentions in the following format:

<table>
<thead>
<tr>
<th>DocID</th>
<th>BeginOffset</th>
<th>EndOffset</th>
<th>PrimaryID</th>
<th>MentionText</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
II. Datasets

Database

We first describe the way we construct our reference knowledge base. Note that participating systems can use the entire Freebase knowledge base, but only entities in our database will be evaluated.

We get our subset by taking a snapshot of Freebase from 9/29/2013, keeping only those entities that have English Wikipedia pages associated with them. We then use the type information in Freebase to further filter unwanted entities. The dataset can be downloaded here: link (MS internal now, will be public later).

In each row, there are three tab-delimited columns:

- FreebaseId
- Name
- WikipediaTitle.

For example, we list all the entities with the name “snow”. Note that the common concept “snow (crystalline water ice)” is not there:

<table>
<thead>
<tr>
<th>FreebaseId</th>
<th>Name</th>
<th>WikipediaTitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>/m/03gr9</td>
<td>&quot;Snow&quot;@en</td>
<td>&quot;/wikipedia/en_title/Snow_200282004_film$0029&quot;</td>
</tr>
<tr>
<td>/m/0c4487</td>
<td>&quot;Snow&quot;@en</td>
<td>&quot;/wikipedia/en_title/Snow_20028Pamuk_novel$0029&quot;</td>
</tr>
<tr>
<td>/m/0cn_w1m</td>
<td>&quot;Snow&quot;@en</td>
<td>&quot;/wikipedia/en_title/Snow_200281963_film$0029&quot;</td>
</tr>
</tbody>
</table>

We keep the entities if they belong to one of the following types:

- "/architecture/building",
- "/architecture/skyscraper",
- "/architecture/structure",
- "/architecture/building",
- "/architecture/house",
- "/architecture/skyscraper",
- "/architecture/tower",
- "/architecture/venue",
- "/aviation/aircraft_model",
- "/award/award",
- "/book/poem_character",
- "/comic_books/character",
- "/comic_strips/character",
- "/cvg/game_character",
- "/fictional_universe/fictional_character",
- "/automotive/model",
- "/automotive/company",
- "/commerce/brand",
- "/comic_books/series",
- "/book/magazine",
- "/book/periodical",
- "/book/written_work",
- "/broadcast/broadcast",
- "/broadcast/radio_station",
- "/broadcast/tv_station",
- "/business/consumer_product",
- "/business/product_line",
- "/business/brand",
- "/commerce/brand",
- "/commerce/consumer_product",
- "/computer/computer",
- "/computer/operating_system",
- "/computer/software",
- "/cvg/computer_videogame",
- "/cvg/developer",
- "/cvg/game_series",
- "/education/university",
- "/film/festival",
- "/film/film",
- "/film/film_series",
- "/internet/website",
- "/internet/website_owner",
- "/location/country",
- "/location/location",
- "/music/musical_group",
- "/organization/organization",
- "/people/deceased_person",
- "/people/person",
- "/sports/pro_athlete",
- "/sports/sports_team",
- "/sports/team",
- "/time/holiday",
- "/time/recurring_event",
- "/tv/network",
- "/tv/tv_network",
- "/tv/program",
- "/tv/tv_series_season"
Files related to short text track

We will provide a public evaluation set composed of a subset of web search queries from past TREC competitions. Each query will be mapped to a set of valid interpretations, based on majority agreement among human judges. Another set of web search queries, annotated using the similar guidelines, will be held out for the final evaluation. Note that the distribution of the public evaluation queries may not be representative of the final evaluation queries.

Files related to long text track

For the long text track, we will use existing ClueWeb annotations (http://lemurproject.org/clueweb09/FACC1/ and http://lemurproject.org/clueweb12/FACC1/), which are produced automatically. The evaluation set and the final testing set will be similar to ClueWeb, which is downloaded from the internet and re-annotated by the human judge.

III. Evaluation Measures

We will only evaluate the entities in the reference database. The predictions of the entities that do not belong to our subset will be discarded. We will have two different evaluation metrics, one for each track.

Short text track

For the short text track, a web search query can legitimately have more than one interpretation. Thus, the ERD systems are expected to generate multiple full query interpretations of non-overlapping linked entity mentions that are semantically compatible with the query text. These interpretations will be evaluated by comparing them to the annotations produced by the majority agreement among 3 human judges.

We evaluate the performance of an ERD system using average F-measure. Specifically, given a query \( q \), with labeled interpretations \( \hat{A} = \{ \hat{E}_1, ..., \hat{E}_m \} \), where each interpretation consists of a set of mentioned entities \( E = \{ e_1, ..., e_L \} \). There is no segmentation and the ordering information stored in \( E \). Note that if there are two mentions referring to the same entity, it will only appear once in the set \( E \). We define the F-measure of a set of hypothesized interpretations \( A = \{ E_1, ..., E_m \} \) as:

\[
F\text{-measure} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
\[
\text{Precision} = \frac{|\hat{A} \cap A|}{|A|} \\
\text{Recall} = \frac{|\hat{A} \cap A|}{|\hat{A}|}
\]

The average F-measure of the evaluation set is simply the unweighted average of the F-measure for each query:

\[
\text{Average F-measure} = \frac{1}{N} \sum_{i=1}^{N} \text{F-measure}(q^i)
\]

Note that a hypothesized interpretation is counted as correct only if it matches all the entities of an interpretation in the reference label set exactly. For simplicity, we do not evaluate the correctness of the entity segmentation in the short text track.

**Long text track**

Let \( \hat{B} = \{\hat{L}_1, \ldots, \hat{L}_n\} \) be a set of reference linked mentions for a document \( d \), where each linked mention \( \hat{L}_j = (\hat{s}_j, \hat{t}_j, \hat{e}_j) \) specifies the begin character offset \( s_j \), the end character offset \( t_j \), and the linked entity id \( e_j \). As the challenge focuses on the disambiguation of entities, we will only evaluate a relaxed correctness of the entity mention boundaries. Specifically, a linked mention \( (s, t, e) \) matches a reference linked mention \( (\hat{s}, \hat{t}, \hat{e}) \) if:

a) \( \hat{e} = e \)

b) \( [\hat{s}, \hat{t}] \) overlaps \( [s, t] \)

There are no overlapping linked mentions in the reference set. The ERD system should produce non-overlapping linked entity mentions as well. For a document \( d^i \), let \( B^i \) represent the output of an ERD system and \( \hat{B}^i \) the reference linked mentions. We define the number of true positives between \( B^i \) and \( \hat{B}^i \) by \( \text{Match}(B^i, \hat{B}^i) \), which returns the maximum number of non-overlapping matches between the hypothesized linked mentions and the reference, where each mention can be mapped at most once. We define the final evaluation metric as the micro-averaged F-measure:

\[
\text{Precision} = \frac{\sum_i \text{Match}(B^i, \hat{B}^i)}{\sum_i |B^i|} \\
\text{Recall} = \frac{\sum_i \text{Match}(B^i, \hat{B}^i)}{\sum_i |\hat{B}^i|} \\
\text{F-measure} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
IV. Participation and Process

The Challenge will be conducted over the internet using the open service format as adopted in the previous two MSR-Bing Challenges. Participants are required to implement their systems as a publicly accessible web service following a REST based protocol defined below and submit their contending entries to the Challenge’s website in the form of web service URI. Upon receiving a submission, service calls to the contending entry will be scheduled for evaluation and metric computation as described above.

The Challenge is conducted in two phases. From the opening of the website to the end of the Challenge D-Date, the website will use a trial dataset to perform system evaluations. This trial phase is mainly for the participants to get familiarized with the tasks and the engineering infrastructure. During this period, each participant can submit multiple systems, each with its own runID as specified below, for evaluations. Participants can direct the Challenge website to post the evaluation results publicly or privately.

Prior to the D-Date, all participants who have entered multiple systems for consideration must choose one as the primary entry to the Challenge. At noon (Pacific Time) on D-Date, the Challenge website will stop receiving new entries and switch to the final test dataset, based on which all the primary entries from the participants will be evaluated with their results publicly posted.

Once all the results of the Challenge are finalized, the Challenge website will resume accepting entries and conduct system evaluations based on the final test set. The website will be maintained indefinitely for the purpose of benchmarking for future researchers.

V. Web Service Interface

Each entry to the Challenge must be a REST based web service that is publicly accessible through HTTP POST with the following request parameters:

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>runID</td>
<td>UTF-8 string</td>
<td>A unique identifier naming a particular evaluation run when the system is submitted to the Challenge website. No white spaces are allowed. The parameter is useful for multiple ERD systems to share the same web service interface.</td>
</tr>
<tr>
<td>TextID</td>
<td>UTF-8 string</td>
<td>The ID of the Text or Query</td>
</tr>
<tr>
<td>Text</td>
<td>UTF-8 string</td>
<td>The text target for ERD.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- For the short text track, this parameter is the raw query that a user has typed into the query box.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- For long text track, this parameter is the document to be evaluated.</td>
</tr>
</tbody>
</table>
Once the ERD task is performed, the web service must respond with a HTTP 200 OK, a response body encode in UTF-8 and the MIME type set to ‘text/plain’. The response body should have the format as follows.

The output of the web service for both long and short tracks are described in the Task Definition.