ERD’14: Entity Recognition and Disambiguation Challenge

SIGIR 2014 Workshop

David Carmel
Yahoo Labs
MATAM
Haifa, 31905 Israel
david.carmel@ymail.com

Ming-Wei Chang
Microsoft Research
One Microsoft Way
Redmond, WA 98052
minchang@microsoft.com

Evgeniy Gabrilovich
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
gabr@google.com

Bo-June (Paul) Hsu
Microsoft Research
One Microsoft Way
Redmond, WA 98052
paulhsu@microsoft.com

Kuansan Wang
Microsoft Research
One Microsoft Way
Redmond, WA 98052
Kuansan.Wang@microsoft.com

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Entity linking, entity disambiguation

1. OVERVIEW

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 output to accommodate some semantic information about the entities recognized in queries and search results. Recent information retrieval studies published in SIGIR have also paid a significant amount of attention to entity-related research. However, techniques for accurate entity recognition and disambiguation are still far from perfect. The motivation of this workshop is to advance the state of the art in 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 known entities in a given collection or knowledge base. Building a good ERD system is challenging because

- entities may appear in different surface forms,
- the context in which a surface form appears often constrains valid entity interpretations, and
- an ambiguous surface form may match multiple entity interpretations, especially in short text.

The challenge has two parallel tracks. In the “long text” track, the challenge focuses on 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. We hope the outcome of this challenge will provide researchers interested in the field with an opportunity to compare different approaches, exchange thoughts, and formulate a shared vision to advance entity research in the future.

The challenge is open to the general public and participants are asked to build their systems as publicly accessible web services using whatever resources at their disposal. The entries to the challenge are submitted in the form of URLs to the participants’ web services. Participants have a period of 3 months to test their systems using development datasets hosted by the challenge website. The final evaluations and the determination of winners are performed on held-out datasets that have similar properties to the development sets. Further details can be found at the ERD challenge website at `http://web-ngram.research.microsoft.com/erd2014`, and detailed rules of the challenge can be found at `http://web-ngram.research.microsoft.com/erd2014/Docs/Detail%20Rules.pdf`. Archived discussions of a dedicated mailing list can be found at `https://groups.google.com/d/forum/erd-2014`.

The ERD challenge is organized as a SIGIR workshop, and every team that submits results to one of the tracks is invited to the workshop to presents its approach. The workshop is co-sponsored by Google and Microsoft.

SIGIR’14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2600734.