ERD'14: Entity Recognition and Disambiguation Challenge

David Carmel
Yahoo! Labs
david.carmel@ymail.com

Ming-Wei Chang
Microsoft Research
minchang@microsoft.com

Evgeniy Gabrilovich
Google
gabr@google.com

Bo-June (Paul) Hsu
Microsoft Research
paulhsu@microsoft.com

Kuansan Wang
Microsoft Research
kuansanw@microsoft.com

Abstract

In this paper we overview the 2014 Entity Recognition and Disambiguation Challenge (ERD'14), which took place from March to June 2014 and was summarized in a dedicated workshop at SIGIR 2014. The main goal of the ERD challenge was to promote research in recognition and disambiguation of entities in unstructured text. Unlike many past entity linking challenges, no mention segmentations were given to the participating systems for a given document. Participants were asked to implement a web service for their system to minimize human involvement during evaluation and to enable measuring the processing times. The challenge has attracted a lot of interest (over 100 teams registered, and 27 of those submitted final results).

In this paper we cover the task definition, issues encountered during annotation, and provide a detailed analysis of all the participating systems. Specifically, we show how we adapted the pooling technique to address the difficulties of gathering annotations for the entity linking task. We also summarize the ERD workshop that followed the challenge, including the oral and poster presentations as well as the invited talks.

1 Introduction

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. In many applications, recognizing and understanding the meanings of entities mentioned in unstructured text is at the core of semantic analysis. The task of entity recognition and disambiguation is to recognize mentions of entities in a given text, disambiguate them, and map them to the entities in a given entity collection or a knowledge base.

To promote research in developing large scale entity recognition and disambiguation algorithms, we organized the Entity Recognition and Disambiguation (ERD) challenge, which

took place from March to June 2014. While some entity linking challenges have been already conducted before, our challenge was different in several respects. First, most previous challenges [18, 12, 11] provided mention segmentations for a given document and focused only on the disambiguation component. In contrast, in the ERD challenge, we evaluated the systems in an end-to-end fashion as mention segmentations were not given to participants. Second, we asked each system to build a web service for their ERD system so that human involvement is minimized during evaluation and the processing time can be measured.

The challenge included two tracks. In the "long-text" track, the challenge targets were pages crawled from the Web; these contain documents that are meant to be easily understandable by humans. In the "short-text" track, on the other hand, the targets were web search queries that are intended for a machine. As a result, the text in the latter track was typically short and often lacked proper punctuation and capitalization.

The aim of this paper is to summarize the challenge results, which were presented at a workshop at the 37th Annual International ACM SIGIR Conference. Some of our major findings are:

- To the best of our knowledge, this is the first time pooling techniques [13, 26, 24] were applied to collect labeled data for the entity linking task. We found that pooling techniques are effective for constructing labeled data, as verifying the correctness of entity assignments is often easier than generating them.
- It is generally easy to create a high precision ERD system. The challenging part is to increase the recall of the system while maintaining good precision.
- To our surprise, the systems that run faster also perform better in the ERD challenge. This is an interesting phenomenon as entity linking systems are often quite complex.

The paper is organized as follows. An overview of the ERD task is presented in Section 2. Section 3 reviews the way we constructed the dataset and describes the pooling procedure. The challenge results are given in Section 4. Section 5 presents some analysis we performed over the challenge log and results. We summarize the workshop in Section 6 and describe the publicly shared resources in Section 7. We conclude in Section 8.

2 Entity Recognition and Disambiguation Challenge

2.1 Task

Given an input text and a knowledge base, the goal of an ERD system is to recognize entity mentions within the text and link (disambiguate) them to entities in the reference knowledge base using their context. In the short-text track, given a search query, we asked participating systems to provide a set of valid entity linking interpretations using all available context. For example, for the query *total recall movie*, the phrase *total recall* should be linked to either the 2012 or the 1990 movie. The phrase *movie* should not be linked as we have excluded the entity *Film* from our knowledge base. Thus, there are two valid interpretations here that should be identified, each containing only one entity.

An ERD system also needs to handle aliases. For example, the query *the governator* should be linked to its eponymous TV program as well as to *Arnold Schwarzenegger*.

In the long-text track, the task is to identify all mentions of entities within a web page that have a reference in a given knowledge base, which 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 to the TV program.

2.2 Data

We constructed our reference knowledge base by taking a snapshot of Freebase from 9/29/2013, keeping only entities with associated English Wikipedia pages. We then used the type information in Freebase to further filter out entities.[1]

. The knowledge-base is publicly available and can be downloaded from `http://web-ngram.research.microsoft.com/erd2014/Docs/entity.tsv`. Although participating systems were allowed to use the entire Freebase knowledge base, only entities in our database were considered for evaluation.

2.3 Evaluation

Short Track: For the short-text track, a web search query can legitimately have more than one interpretation. Thus, a good ERD system is expected to generate multiple query interpretations of non-overlapping linked entity mentions that are semantically compatible with the query text. Such interpretations are evaluated by comparing them to the annotations produced by majority agreement among 3 human judges using average F-measure.

Specifically, given a query q, with labeled interpretations $\hat{A} = \{\hat{E}_1, \ldots, \hat{E}_n\}$, where each interpretation consists of a set of mentioned entities $E = \{e_1, \ldots, e_l\}$ without segmentation and ordering information. Thus, 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 $A = \{E_1, \ldots, E_m\}$ as follows:

$$F\text{-}measure = 2(Precision \times Recall)/(Precision + Recall)$$

$$Precision = |\hat{A} \cap A|/(|A|)$$

$$Recall = |\hat{A} \cap A|/(|\hat{A}|)$$

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

$$Average\ F\text{-}measure = \frac{1}{N}\sum_{i=1}^{N} F\text{-}measure(q_i)$$

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

[1]We used a set of "good" entity types and a set of "bad" entity types to construct the entity snapshot. The entities that have bad entity types were thrown away, while the entities that have the good entity types survived in our entity snapshot. The good type set includes types such as "/location/location", "/organization/organization" and "/people/person". The bad type set includes types like "/book/magazine_genre", "/book/newspaper_edition_type" and "/broadcast/tv_signal_type".

Long 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 beginning character offset s_j, the end character offset t_j, and the linked entity id e_j. Since mention segmentation is often ambiguous, the challenge focused on the disambiguation of entities, and hence we only evaluated 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:

$$\begin{cases} \hat{e} = e, \text{and} \\ [\hat{s}, \hat{t}) \text{ overlaps with} [s, t) \end{cases}$$

Since 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 denote the number of true positives between B^i and $\hat{B}^i$ by $Match(B^i, \hat{B}^i)$, which returns the maximum number of non-overlapping matches between the hypothesized linked mentions and the reference set, where each mention can be mapped at most once. We define the final evaluation metric as the micro-averaged F-measure:

$$Precision = (\sum_i Match(B^i, \hat{B}^i))/(\sum_i |B^i|)$$

$$Recall = (\sum_i Match(B^i, \hat{B}^i))/(\sum_i |\hat{B}^i|)$$

$$F\text{-}measure = 2(Precision \times Recall)/(Precision + Recall)$$

Evaluation Protocol: In the ERD challenge, each team hosted a web service for their ERD system. When a team issued an evaluation request to the evaluation server (hosted by the ERD organizers), the evaluation server sent a set of *unlabeled* texts to the team's web service to retrieve the corresponding predictions. The returned results were then evaluated, and the evaluation score was posted on the challenge leaderboard.

3 Dataset Construction

3.1 Document Collection Process

There are two main sets of documents for each track of the ERD challenge. We refer to the document collection used during the development period (from March 25th to June 10th) as the development set. A disjoint test set collection was used during the test period (from June 10th to June 20th). As evaluation was performed on the evaluation server, the labeled data for both development and test sets were not released to the participants.[2]

For the short-text track, we sampled 500 queries from a query log of a commercial search engine to form a development set and 500 queries for the test set. The average query length was 4 words per query.

For the long-text track, we sampled 100 web pages for the development set and another 100 web pages for the test set. We stripped out all HTML tags from the web pages and

[2]To clarify the task definition, we released some labeled data from the development set.

applied various heuristics to extract the main content from each document. Specifically, boilerplate content from header and side panes were removed. Among all documents, 50% were sampled from general web pages; the remaining 50% were news articles from `msn.com`.

3.2 Annotation Guideline

Defining a clear annotation guideline is an important yet challenging step in designing the ERD challenge. Recall that the task is to identity mentions of entities from a predefined entity set. However, exactly what constitutes an entity mention is often ambiguous and highly debated, with seemingly inconsequential differences significantly affecting the system performance. In this challenge, we employed the following annotation guidelines:

- Use the longest span for an entity.

  I live in Redmond, WA should be annotated as *I live in [Redmond, WA]*, where we use brackets to mark the boundary of a mention. If the city *Redmond, WA* is not in our database, the ERD system should produce *I live in Redmond, [WA]*, mapping *WA* to the state of Washington.

- Annotate named entities only.

  In the ERD challenge, we mainly focus on proper noun entities. Therefore, *Kobe Bryant's wife* should be annotated as *[Kobe Bryant]'s wife*.

- No overlapping entities are allowed.

3.3 Pooling

Generating entity mention labels is difficult and time-consuming for human judges, given that annotators often need to verify various possible entity assignments. Labeling entities often requires the annotators to be knowledgeable about the entities as well.

For the long-text track, there are potentially hundreds of candidate mentions in a document, each potentially linking to multiple entities. The fact that candidate mentions may overlap with one another further increases the difficulty (note, however, that the labeled entities never overlapped).

Labeling short-text track queries is also time consuming as we requested short-text ERD systems to produce all possible entity interpretations. For example, for the query *robert walker actor*, the annotator needs to list *all* actor entities in Wikipedia with the name *robert walker*.

Interestingly, *verifying* the correctness of an entity-mention pair is relatively easy, if the entity information is clearly presented in front of the judge. Thus, we can easily compute the precision of ERD systems because it only involves verifying hypothesized entity-mention pairs.

Calculating recall, on the other hand, is more challenging. To overcome the difficulties, we applied the *pooling* technique [26] frequently used in the information retrieval community to supplement our initial labels. As participants submit their system results, we identify previously unlabeled mention-entity pairs and verify their correctness via crowdsourcing.

In the crowdsourcing step, a judge is presented with the text, the target mention, and information about the candidate entity. The job of the judge is to answer a yes-no question: does the entity represent the meaning of the mention in context. Each mention-entity pair

is judged by at least three judges, with the majority consensus used in the subsequent evaluation.

After the initial crowdsourcing judgments, there may still be inconsistencies in the annotations. For example, given that each judge only sees one mention-entity pair at a time, there may be overlapping mentions in the resulting annotations. Judges may also accept non-proper noun entity mentions. Therefore, in the conflict resolution and post editing step, a professional annotator reviewed all the annotations and corrected the mistakes made in the previous steps. Based on our analysis, the labeled data generated by these steps are quite good.

Specifically, our pooling procedure consists of:

1. Collecting distinct annotations from all participating ERD systems.
2. Verifying the correctness of the annotations via crowdsourcing.
3. Performing conflict resolution and post editing of the crowdsourced labels.
4. Evaluating the performance of each ERD system against the final labeled data.

Periodically, during the development period, we published a new version of the development set on the leaderboard that includes the additional labels from pooling, and sent out a notification so participants could resubmit their results for evaluation against the new version. The same pooling technique was applied to obtain the labels for the final test set at the end of the competition.

To the best of our knowledge, this is the first time pooling has been used in information extraction tasks. It is important to note that pooling does not guarantee the identification of all mentioned entities in the document, as some entities may be missed by all systems. However, system ranking still represents the relative performance [26].

4 Challenge Results

Participants Overview: About 100 teams registered to ERD 2014. In total, there were 27 teams, from both academia and industry, who submitted their final results. Among the 27 teams, 6 teams submitted to both tracks, 9 teams only to the long-text track, and 12 teams only to the short-text track. Participating teams who submitted final reports are listed in Table 1. Note that in addition to the 18 teams listed in Table 1, 9 teams submitted the final results but did not submit their reports.[3]

Results: The results for the short and long-text tracks are presented in Figure 1a and Figure 1b, respectively. The winner of the short-text track is the SMAPH team [3] followed by NTUNLP [2] and Seznam Research [7]. The winner of the long-text track is the MS_MLI team [4], followed by MLNS [17] and Seznam Research [7].

It is worthwhile noting that the final evaluation took into account the processing time. In the long-text track, document processing time was limited to 60 seconds. In the short-text track the processing time was restricted to 20 seconds per query.

[3]They are C3 (Both), InriaBerlin (Short), XI-lab (Short), UIUC-GSLIS (Short), HITS (Long), MLNS2 (Long), NERD (Long), Acube Lab 2 (Long) and Ryiko (Long).

Team Name	Report	Institution	Track
ExPoSe	[21]	University of Amsterdam, University of Tehran	Both
NTUNLP	[2]	National Taiwan University	Both
Seznam Research	[7]	Seznam Research	Both
UBC	[1]	University of the Basque Country	Both
WebSAIL	[20]	Northwestern University	Both
CI_ERD	[15]	IIT Bombay, Stanford University	Short
clues_ERD	[6]	CMU	Short
Magnetic_IISAS	[16]	Magnetic Media Online, Slovak Academy of Sciences	Short
NTNU-UiS	[10]	Norwegian University of Science and Technology, University of Stavanger	Short
SIEL@ERD	[23]	International Institute of Information Technology, Hyderabad	Short
SMAPH	[3]	University of Pisa, Google, University of Munich	Short
TALP-UPC	[19]	TALP Research Center	Short
UvA	[9]	University of Amsterdam	Short
whu_ir	[25]	Wuhan University	Short
Acube Lab	[22]	University of Pisa	Long
MLNS	[17]	Dalhousie University	Long
Neofonie	[14]	Neofonie GmbH	Long
MS_MLI	[4]	Microsoft Research	Long

Table 1: List of the teams who submitted both final results and reports. In addition to these 18 teams, 9 teams submitted the final results but did not submit their reports.

5 Analysis

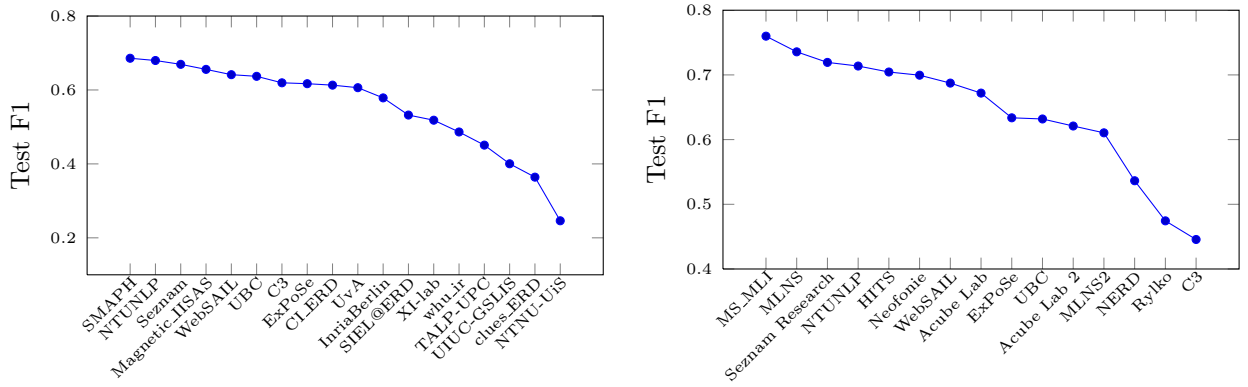
In this section, we provide some basic analyses based on the submission log and the results of the participating ERD systems. Specifically, we address the following research topics:

- The relative performance of the ERD systems
- The relationship between the number of submissions during the development period and the system performance
- Is the processing time correlated with the final performance?
- How diverse are the results produced by different ERD systems?

Precision and Recall: Before presenting the system results, note that our evaluation metric for the short-text track is quite different than the evaluation metric for the long-text track (cf. Section 2.3). For the short-text track, the precision and recall are calculated to handle multiple interpretations per query[4], making it trickier to report an overall precision and recall.

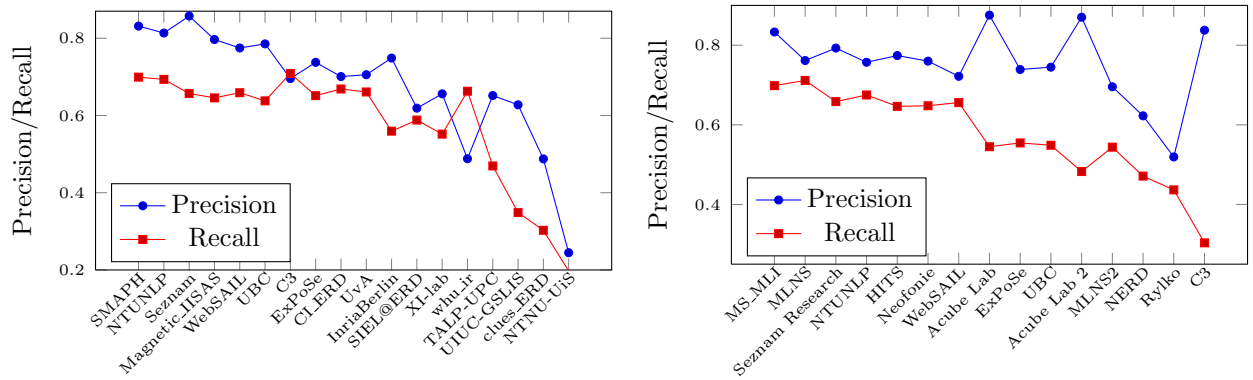
For analysis purposes, we recalculated the precision and recall for the short-text track at the entity level, the same way as they were calculated in the long-text track. Specifically, an entity annotation is considered correct if it appears in the gold labeled data. By calculating the precision and recall this way, entities belonging to different interpretations were mixed together. For the long-text track, we use the same calculation as defined in Section 2.3.

[4]For instance, *Robert Walker* may be linked to multiple entities given that there are multiple Wikipedia entities named Robert Walker.



(a) Short Text Track (b) Long Text Track

Figure 1: Final results of the ERD challenge.



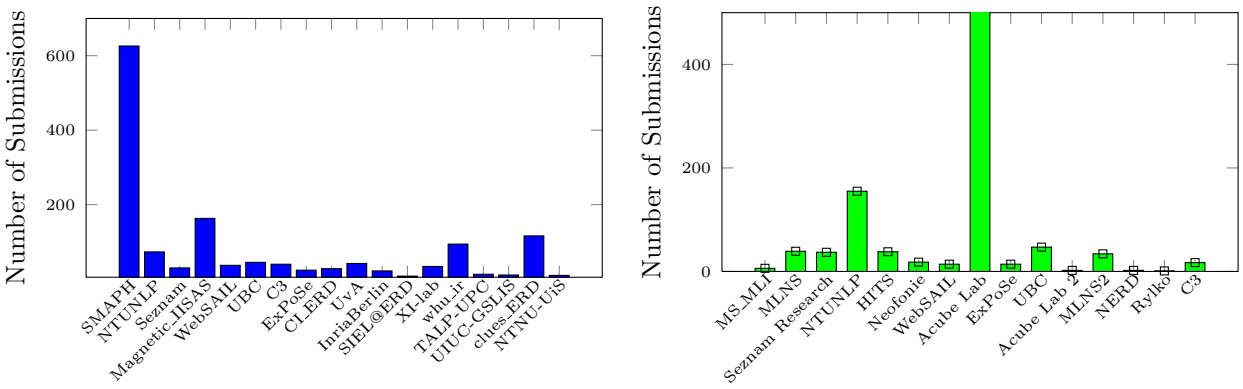
(a) Short Text Track (b) Long Text Track

Figure 2: Precision and recall of participant teams in both tracks. Note that for the short-text track, we used different definitions of precision and recall than those defined in Section 2.3.

The precision and recall results are presented in Figure 2. First, note that the precision is usually significantly higher than the recall for almost all systems. This holds for both tracks. Second, while some ERD systems in the long text track have very good precision, they do not achieve the top overall performance. It is probably easy to spot some unambiguous entity mentions in the text with high precision. However, it is challenging to also paying attention to recall while keeping a good precision.

Another interesting discovery is that while the short-text track was designed to encourage systems to handle multiple interpretations, many teams output only the top interpretation. In our final testing data, we identified 530 entity assignments among the 500 queries; 25 of the mentions have multiple interpretations. However, among the top performing systems, only Seznam Research handled multiple interpretations while SMAPH and NTUNLP only generated the top interpretation. Therefore, it seems to be difficult to build a robust system that handles multiple interpretations without sacrificing precision.

Number of Submissions: There was no limitation on the number of submissions per system during the development period. Hence, teams could submit results as many times as



(a) Short Text Track

(b) Long Text Track

Figure 3: The number of submissions of each team during the development period.

they wished to be evaluated against the development set. It is interesting to analyze whether there is any correlation between the number of system's submissions during the development period and its final performance on the test set.

Note that we reset the leaderboard and the development set several times during the development period (see Section 3). For this analysis, we only used the statistics from the final version of the development dataset, which was released 10 days before the testing period.

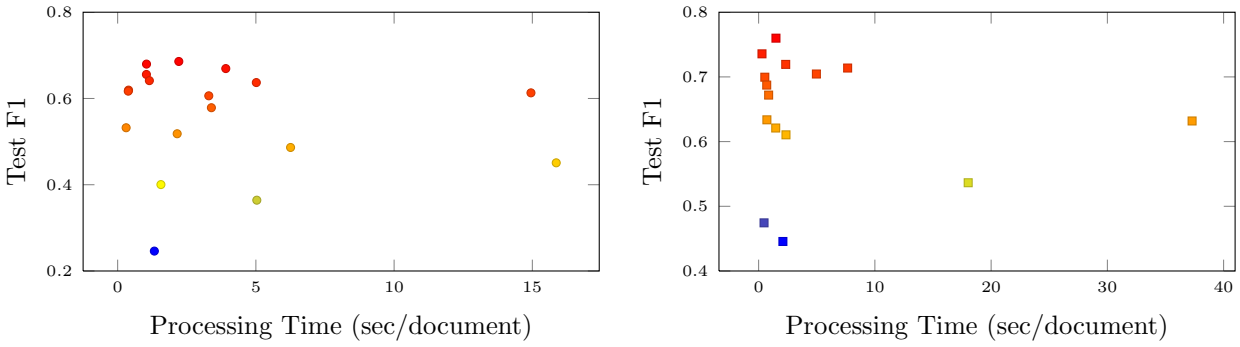
The statistics of the ERD systems submissions are presented in Figure 3. First, we can see that the ERD teams were very active, with 3152 submissions across all teams within 10 days. Second, a few teams used the submission systems to automatically tune their hyper parameters on the development set. Finally, the number of submissions and the final performance (F1) demonstrated a surprisingly weak correlation of only 0.31. The correlation coefficient on the long-text track is merely 0.13. The release of some labeled data for the development set in the long-text track may explain the particularly weak correlation, as teams may have chosen to perform the analysis offline.

Processing Time: In the ERD challenge, the processing time was also evaluated by recording the average number of seconds for an ERD system to generate prediction results given a text. Note that the processing time also includes the networking overhead as the data are transmitted through the web.

Figure 4 shows the relationship between the processing time and the final performance. Given that ERD systems are often quite complicated, we expected that the systems that utilize more time would perform better. However, surprisingly, the systems that perform better are also often faster. This may imply that currently state-of-the-art performance may be achieved without the usage of time-consuming natural language processing components.

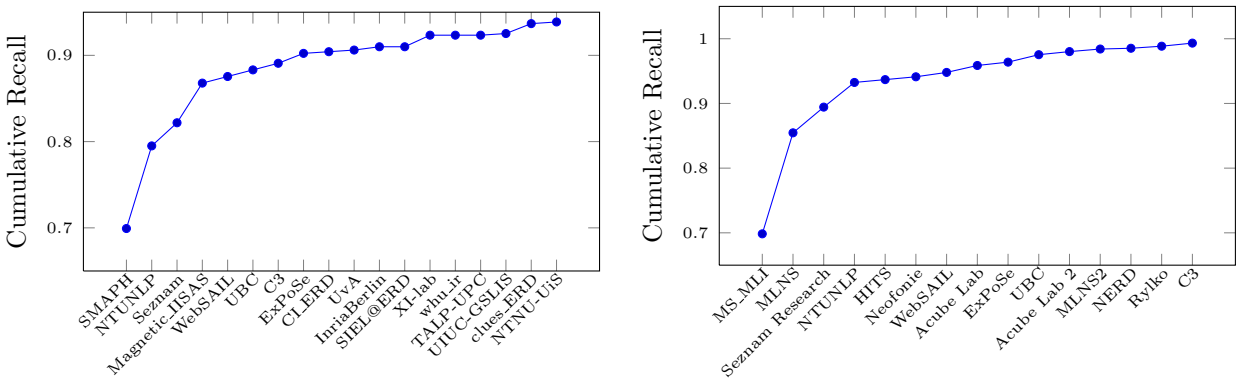
System Diversity: The labeled data for system evaluation was generated via pooling. Therefore, the diversity among systems results is quite important. In this subsection we analyze how many systems are needed to form an effective pool.

To evaluate the diversity among systems, we used the *Cumulative Recall* metric, which is calculated as follows:



(a) Short Text Track (b) Long Text Track

Figure 4: The plot between the processing speed and the final performance.



(a) Short Text Track (b) Long Text Track

Figure 5: The cumulative recall of the ERD systems.

1. We rank the ERD teams by their final performance. We initialize a common set to be the empty set.
2. We traverse over the teams by increasing rank. At each stage, we add the results of the system into the common set.
3. We report the recall of the common set until the results of all teams have been visited.

The results of the cumulative recall are presented in Figure 5.[5] From the figure one can observe that the ERD systems did produce very diverse results. In both tracks, the top two systems in the pool increased the cumulative recall by 10% over the top system. Moreover, four systems in the pool already covered about 90% of the entities in both tracks.

6 Workshop Report

The ERD challenge was summarized in a one-day workshop during the SIGIR'14 conference. In the following we overview the workshop talks and presentations.

[5]Since for the short-text track, we have some additional seed annotations, the final cumulative recall does not reach 100%.

6.1 Invited Talk

The workshop began with a keynote talk given by James Zhang from the Knowledge Engineering group at Bloomberg. Zhang focused on the research directions that Bloomberg Labs is pursuing in the area of ERD in the financial and business domains. He discussed several challenges with a real-world application of predicting changes in stock prices in response to a news report. Such an application demands highly precise named entity recognition and disambiguation, in order to reduce false alerts based on incorrect entity detection. Moreover, the scale and diversity of the data to be analyzed is challenging, as customers can subscribe to multiple sources of information – national news, blogs, and many other social media sites.

ERD from multilingual data is also challenging as most non-English languages lack appropriate tagged data for training. One of the directions they experimented with is using statistical machine translation approaches for aligning named entities extracted from the multilingual data using several well-trained ERD tools. Zhang also discussed the static nature of the knowledge bases used by the ERD tools and the difficulty they pose. As certain entities have very short half-lives in the financial market, new emerging entities which are typically the most interesting ones cannot be detected while using static, infrequently updated knowledge bases.

6.2 Paper Presentation

The keynote talk was followed by a summary of the ERD challenge given by the organizers, and the main results achieved in the short-text track and the long-text track. The challenge winners were announced and prizes where delivered to the winning teams. Then, seven invited teams presented their ERD systems, their experimental results, and the lessons learned from the challenge.

Yen-Pin Chiu from NTU represented the NTUNLP team [2], which created an ERD system that is based on a dictionary of Freebase entities. By scanning the given text while using the longest match strategy of dictionary strings, candidate entities were detected within the text. Some experiments with selection methods of best candidates using several external resources such as DBpedia Spotlight [5] and Tagme [8] were reported. The NTUNLP team placed second in the short track and fourth in the long track.

The second speaker was Jaap Kamps from the University of Amsterdam who presented the Expose system [21] for ERD. The focus of this team was to improve on the results of an open source ERD system, namely DBpedia Spotlight. One of the main drawbacks of this system is the identification of cases where a name does not refer to any known entity. To improve this so-called NIL[6]-detection, a classifier was designed and trained to automatically classify DBpedia Spotlight's output entities as NIL or not-NIL. The analysis of this system on the challenge's datasets showed that the proposed approach successfully improved the accuracy of the baseline system.

Marco Cornolti from the University of Pisa presented the SMAPH system [3], which ranked first in the short-text track. The SMAPH system implemented a pipeline of four main steps: (1) Fetching: fetch the search results returned by a general purpose search engine (Bing) given the query to be annotated; (2) Spotting: search result snippets are parsed to identify candidate entities by looking at the boldfaced parts of the search snippets; (3) Candidate generation: candidate entities are generated in two ways: from the Wikipedia

[6]NIL - Not In Lexicon

pages occurring in the search results, and from the mentions identified in the spotting step as input; (4) Pruning: a binary SVM classifier was used to decide which entities to keep or discard in order to generate the final annotation set for the query.

Marek Lipczak from Dalhousie University presented the Tulip system [17], which participated in the long-text track. One of the key features of Tulip is the novel representation of the processed document and the spotted entity candidates as vectors of Wikipedia categories. An importance score for each entity's category is calculated by unifying information from 120 language versions of Wikipedia, where each language version acts like an independent witness of relations. A document is represented by generating a centroid based on its core identified entities. This allows the system to quickly estimate the coherence of each candidate entity with the other document candidates. Tulip ranked second in the long track. Additionally, the system was the fastest among all participants.

Francesco Piccinno from the University of Pisa presented WAT [22], a successor of the classic Tagme system. This extension included re-designing all Tagme components, namely, the spotter, the disambiguator, and the pruner. With respect to Tagme, the improvement in WAT performance ranged from 1% to 9% on the development dataset. One of the main conclusions from the experiments with WAT was that while many ERD systems focused on improving disambiguation, the spotter and the pruner modules actually are responsible for introducing many of the false positives. By redesigning these two components, WAT was able to significantly improve the quality of the results.

Finally, Jiri Materna from Seznam Research described their system [7], which participated in both tracks. Their entity disambiguation process is based on co-occurrence analysis using a graph of candidate entities, with links between entities extracted from Wikipedia articles. The speaker mentioned that although the system does not introduce any ground-breaking theoretical results, the actual performance of the system, motivated by simple, logical heuristics, and without heavy parameter tuning and machine learning, performed very well and ranked third in both tracks.

6.3 Poster Sections and Discussions

The oral talks were followed by a poster session. All ERD participants were invited to present their work during the poster session and six teams participated in presenting their work. The session provided a great opportunity for the participants to share their experience and get feedback from others. In addition, it stimulated a lot of discussions about the ERD challenge in general, and more specifically how to proceed in the future.

7 Public Shared Resources

After the ERD competition, we released and shared several resources to the public. All of the resources can be accessed through the official challenge website: `http://web-ngram.research.microsoft.com/erd2014`.

- Since the organizers still maintain the evaluation service, everyone can build a web service and have their ERD system evaluated under the official settings.
- Several labeled datasets can be downloaded from `http://web-ngram.research.microsoft.com/erd2014/Datasets.aspx`.

- Some participants have made their servers public. The server information is available at `http://1drv.ms/1sGu1EP`.
- The reports of most ERD systems are available online and linked from the challenge website above.

8 Conclusion

In this challenge we proposed a new way to collect labeled data via the pooling technique and compared various aspects of different entity linking systems. We are pleasantly surprised at the diversity of approaches among the teams. For example, the SMAPH [3] team chose to use a search engine as a component in their entity linking system. The NTUNLP [2] team performed system combination. The MLNS [17] system, one of the best performing system in the long-text track, focused largely on the mention detection problem.

However, there is still room for improvement in our data construction and pooling procedure. First, some important entities are not included in our entity snapshot while some common noun entities are included. Therefore, one interesting research direction is to determine the important entities that specific application scenarios care about. We also spotted some mistakes made by the professional annotators in the conflict resolution step. Thus, another important research direction is to study the effect of pooling and propose a better procedure to further improve the quality of the labeled data.

As the evaluation service is still active, teams have already improved their systems after the ERD workshop. At the time of writing of this report, teams have already achieved results that exceed the best results from the official testing run on both tracks. We hope that through our release of data and resources, we will promote the research on entity recognition and disambiguation.

Acknowledgment The organizers sincerely thank Siyu Liu and Charles Huang for their help and devotion to the creation of the dataset and maintenance of the evaluation server.

References

[1] Ander Barrena, Aitor Soroa, and Eneko Agirre. UBC Entity Recognition and Disambiguation at ERD 2014. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[2] Yen-Pin Chiu, Yong-Siang Shih, Yang-Yin Lee, Chih-Chieh Shao, Ming-Lun Cai, Sheng-Lun Wei, and Hsin-Hsi Chen. NTUNLP Approaches to Recognizing and Disambiguating Entities in Long and Short Text in the 2014 ERD Challenge. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[3] Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita, Hinrich Schütze, and Stefan Rüd. The SMAPH System for Query Entity Recognition and Disambiguation. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[4] Silviu-Petru Cucerzan. Name Entities Made Obvious. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[5] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[6] Bhavana Dalvi, Chenyan Xiong, and Jamie Callan. A Language Modeling Approach to Entity Recognition and Disambiguation for Search Queries. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[7] Alan Eckhardt, Juraj Hreško, Jan Procházka, and Otakar Smrž. Entity Recognition Based on the Co-occurrence Graph and Entity Probability. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[8] P. Ferragina and U. Scaiella. TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM*, 2010.

[9] David Graus, Daan Odijk, Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Semanticizing search engine queries: the University of Amsterdam at the ERD 2014 challenge. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[10] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. A Greedy Algorithm for Finding Sets of Entity Linking Interpretations in Queries. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[11] H. Ji, R. Grishman, and Dang. Overview of the tac 2011 knowledge base population track. In *Proceedings of the TAC 2011 Workshop*, 2011.

[12] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, 2010.

[13] SPARCK JONES. Report on the need for and provision of an" ideal" information retrieval test collection. 1975.

[14] Steffen Kemmerer, Benjamin Großmann, Christina Müller, Peter Adolphs, and Heiko Ehrig. The Neofonie NERD System at the ERD Challenge. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[15] Ashish Kulkarni, Kanika Agarwal, Pararth Shah, Sunny Raj Rathod, and Ganesh Ramakrishnan. System for Collective Entity Disambiguation. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[16] Michal Laclavúk, Marek Ciglan, Alex Dorman, Štefan Dlugolinský, Sam Steingold, and Martin Šeleng. A Search Based Approach to Entity Recognition: Magnetic and IISAS team at ERD Challenge Short-text track. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[17] Marek Lipczak, Arash Koushkestani, and Evangelos Milios. Tulip: Lightweight Entity Recognition and Disambiguation Using Wikipedia-Based Topic Centroids. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[18] Paul McNamee and Hoa Trang Dang. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC 2009)*, volume 17, pages 111–113, 2009.

[19] Ali Naderi, Horacio Rodriguez, and Jordi Turmo. The TALP Participation at ERD 2014. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[20] Thanapon Noraset, Chandra Sekhar Bhagavatula, and Doug Downey. WebSAIL Wikifier at ERD 2014 Both tracks. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[21] Alexander Olieman, Hosein Azarbonyad, Mostafa Dehghani, Jaap Kamps, and Maarten Marx. Entity Linking by Focusing DBpedia Candidate Entities. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[22] Francesco Piccinno and Paolo Ferragina. From TagME to WAT: a new Entity Annotator. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[23] Priya Radhakrishnan, Romil Bansal, Manish Gupta, and Vasudeva Varma. Exploiting Wikipedia Inlinks for Linking Entities in Query. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[24] Ellen M Voorhees and Donna Harman. Overview of trec 2001. In *Trec*, 2001.

[25] Chuan Wu, Wei Lu, and Pengcheng Zhou. An Optimization Framework for Entity Recognition and Disambiguation Short-text track. In *ERD'14: Entity Recognition and Disambiguation Challenge*. ACM, 2014.

[26] J. Zobel. How Reliable Are the Results of Large-scale Information Retrieval Experiments? In *SIGIR*, pages 307–314, New York, NY, USA, 1998. ACM.