Name Entities Made Obvious: The Participation in the ERD 2014 Evaluation

Silviu Cucerzan
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

ABSTRACT
The paper describes NEMO, a system for extracting entity mentions from text and linking them to Wikipedia (and Freebase), which participated in the ERD 2014 challenge. The model employed by the system allows a seamless use of traditional priors and lexical features in conjunction with various types of latent features, which are computed based on the attributes associated with all extractions of entity mentions from an input text and their possible linkage to Wikipedia. Additionally, it allows a unified approach for handling both features computed globally, at document level, and features computed based on the local context, such as syntactic patterns, of each hypothesized entity mention. The model is trained on a large dataset derived from Wikipedia, and achieves state-of-the-art results on the datasets in the ERD evaluation without employing explicitly ERD-specific training data.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing.*

General Terms
Algorithms, Experimentation.

Keywords
Entity recognition and disambiguation.

1. TASK AND FRAMEWORK
The previous large-scale evaluations for entity disambiguation organized by NIST and LDC in the Text Analysis Conference (TAC) have focused on the disambiguation of one entity mention at a given offset in a document [5]. For most of the data points in the TAC sets, the boundaries indicated by the provided offset and length are the exact boundaries of the entity mention targeted for evaluation. Rarely, those boundaries need corrections, because the entity mention in the given text is a superstring (e.g. "Jane" → "`Thomas Jane`") or substring (e.g., "German ARD" → "`ARD`") of the given target string. However, for all data points in the development and test sets, the task setting guarantees the existence of an entity mention that needs be disambiguated. One of the additional challenges in the TAC evaluations has been that a large number of the targeted mentions (about half) refer to entities that are not in the employed knowledge base; systems are required to mark them as NIL, and cluster together the mentions that refer to the same *unknown entity*.

By contrast, in the ERD evaluation, only text documents are provided as input, without any specified target mention. Systems are asked to analyze each document entirely and extract the entities occurring in the document (an example is shown in Figure 1). While TAC gives an important role to NILs, ERD focuses on the extraction of mentions of entities in a given knowledge base and only those entities. The task is made more difficult by the requirement of extracting the mention of a *known entity* even when the mention is part of a longer, but unknown entity, as shown in Figure 2 for the Wikipedia entity "Politics of Australia" with the mention "`Australian politics`". Because this entity is not in the target knowledge base, the system is required to extract "`Australian`" as a mention of the entity "Australia" instead of assigning a NIL id to the full mention.

Another difference between TAC and ERD is the quality and structure of the input text. While TAC has employs documents annotated using a standard XML markup, with clear headers, titles, and consistent formatting, the ERD challenge targets somewhat noisy text extracted from Web pages, including eBay listings and other product listings from commercial Websites. The presence of footers, copyright notices, tables with capitalized entries, and other such elements adds a new challenge to the text processing and the identification of entity mentions.

2. SYSTEM ARCHITECTURE
NEMO (acronym for Named Entities Made Obvious) is a system that preserves the general architecture of the MSR system that participated in the TAC 2013 evaluation with the id MS_MLI [4]. Essentially, the model follows the paradigm that the best evidence for identifying and disambiguating an entity in a context pertains to the properties of and relationships to the set of co-occurring entities in that context.

The remaining of this section describes the content of NEMO's knowledge base, as derived from Wikipedia, the text analysis process, and the disambiguation process employed by the system. The customizations done for ERD are described in Section 4.

2.1 The Knowledge Base
The knowledge base is comprised of three main components: the entity repository, the known entity forms, with priors for mapping them to entities, and the linguistic resources. For the ERD 2014 challenge, all these components are derived from the Wikipedia dump file from August 5, 2013 and the Wikipedia to Freebase mapping file provided by the ERD organizers.

ERD'14, July 11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2633211.2634360

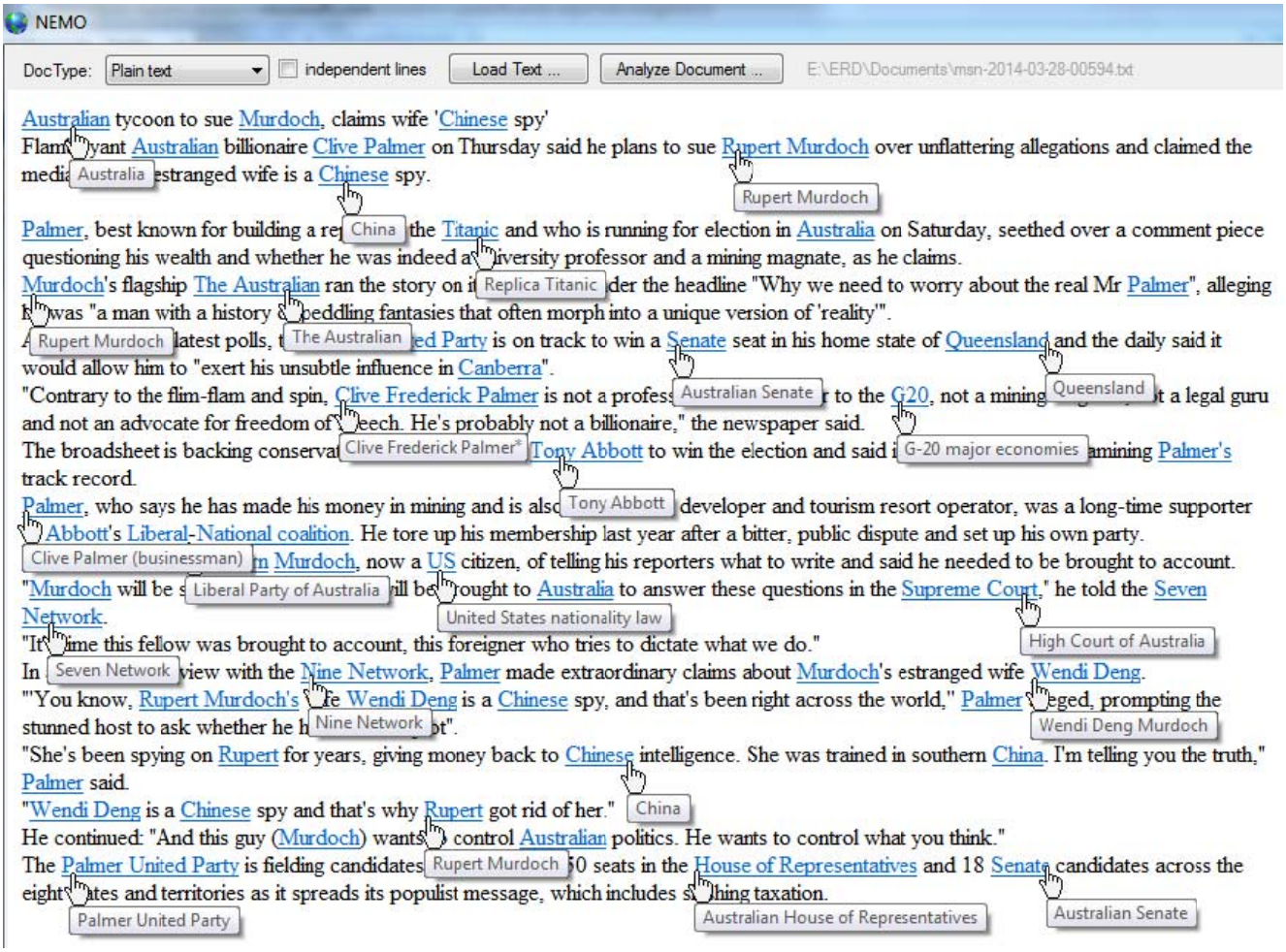

Figure 1. Screenshot of a an analyzed document from the ERD 2014 development set, with superimposed Wikipedia entity names (i.e., disambiguations) as predicted for several of the surface forms extracted by the system.

2.1.1 Entities

For each entity, the system employs the following type of attributes: entity types, topics, triggers, contexts, geo-coordinates, and Freebase IDs. We provide below concise definitions for these attributes and the manner in which their values are obtained.

The *topics* that an entity belongs to are derived from Wikipedia categories, Wikipedia list pages, Wikipedia enumerations of interlinks, Wikipedia tables, and Wikipedia templates. In general, topics group entities that are similar, such as football teams, football players, and football stadiums/venues. Figure 2 shows the number of topics derived from each Wikipedia source, while Figure 4 plots the number of entities in the knowledge base that have a certain number of topics associated to them.

The *triggers* of an entity are defined as other entities that are in close relationship with it; a close relationship is hypothesized between two entities whenever there exists bidirectional linking between their associated Wikipedia pages.

Note that both topics and triggers do not appear explicitly in the text of an input document; they are surfaced only when the presence of entity mentions is hypothesized.

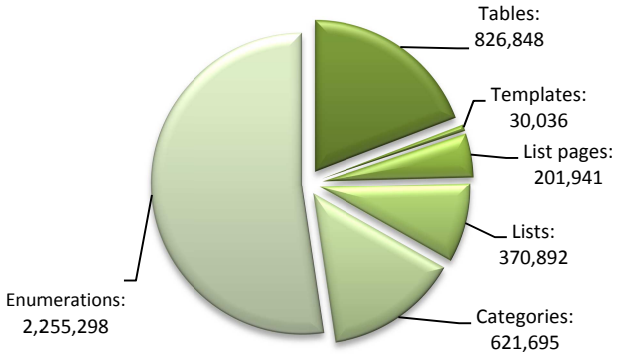

Figure 2. The number of different types of topics in the knowledge base as derived from the Wikipedia collection (August 5, 2013).

By contrast, the *contexts* of an entity are terms that are expected to co-occur in text with the entity mention. They are derived from Wikipedia titles (such as parentheticals), Wikipedia infoboxes, and appositives found in the Wikipedia text.

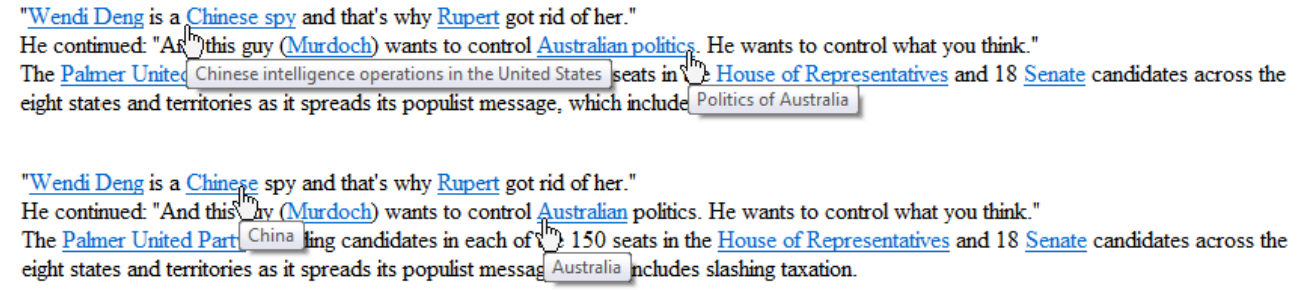

Figure 3. Example of text analyzed by the two versions of the NEMO system: (top) regular analysis; (bottom) ERD-customized analysis, in which the system is aware of which targeted entity set and selects surface form boundaries for which the disambiguation is in the targeted set. Note that the regular system extracts the surface forms `Chinese spy` and `Australian politics`, while the ERD version extracts instead `Chinese` and `Australian`.

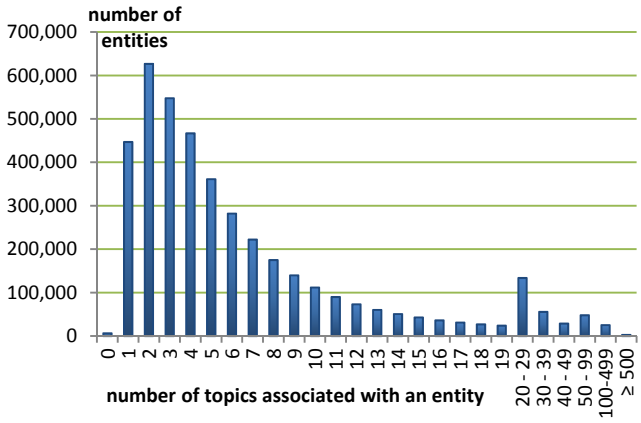

Figure 4. Histogram showing the number of entities in the knowledge base derived from Wikipedia that have various numbers of topics associated with them.

The *types* of entities for the ERD evaluation are derived from the Freebase types associated with the target entities. Therefore, multiple types can be associated to each entity.

Geo-coordinates (latitude and longitude) are extracted from the Wikipedia text for those pages that have such information encoded by using the standard markup. They are normalized in the extraction process, so that distances between any pair of geo-coordinates can be easily computed at runtime.

Similarly, the *Freebase IDs* are associated to entities based on the Wikipedia to Freebase mapping file provided for the ERD challenge. These IDs are used only to determine whether an entity in the system's knowledge base is present in the ERD's target collection.

2.1.2 Entity Forms

Entity forms are defined as strings that can be used to mention entities in text, such as "`Washington`" for the any of the entities "George Washington", "Washington, D.C.", "Washington (state)", and "Federal government of the United States".[1]

[1] In fact, the particular entity form "`Washington`" has no fewer than 386 associated entities in the knowledge base employed for the ERD 2014 challenge.

The entity forms are extracted from multiple Wikipedia sources, including page titles, anchor text of Wikipedia interlinks, Wikipedia infobox fields such as nicknames, Wikipedia redirects, and bolded terms in the Wikipedia text.

For each entity form, the knowledge base stores a probabilistic distribution over all entities that can be mentioned by the form. The values in the distribution are derived based on Wikipedia interlinks statistics i.e., the number of times a form is used as the anchor text for interlinking to the Wikipedia page of each entity.

2.1.3 Linguistic Resources

The system also employs a set of linguistic resources derived from the Wikipedia collection, including *word capitalization statistics* (how many times a word appears capitalized versus lowercased inside Wikipedia sentences), lists of first names and last names (based on histograms for the first and last word in the canonical form of Wikipedia entities of the type person), name normalization (which first name forms can be used interchangeably), distributions over entity types for immediate left and right contexts, etc. These resources are used mainly in the identification of surface forms in text and for mapping identified surface forms to possible entities (by combining entity mapping information for variants of names). The exception to this usage is represented by the distributions over entity types, which are derived by employing the Wikipedia inter-linkage in conjunction with the types associated to the entities in the collection. These distributions are used directly to compute two local-context features in the disambiguation process.

2.2 Text Analysis

For any input text, the system first normalizes the text with respect to spacing and other text delimiters. It then breaks the text into sentences by using word capitalization statistics from the input text and from the Wikipedia collection. Each sentence is then analyzed by a component that hypothesizes mentions of entities in the text. Following the terminology from [2], we refer to the hypothesized entity mentions as *surface forms*.

The identification of surface forms in the text is achieved by a combination of rules based on the linguistic resources and the entity forms derived from Wikipedia.

In some cases, multiple surface forms that are known as entity forms can be identified in the same place in text. For example, in the sentence "And this guy (Murdoch) wants to control Australian politics." from the example text shown in Figure 1, both "`Australian`" and "`Australian politics`" are known

entity forms. As described in [3], the system employs postpones making a decision on selecting one of these surface forms until the disambiguation stage, in which information about the entity candidates corresponding to all surface forms in the document becomes available. To do this, the system builds on the fly a *composite surface form* with an entity mapping vector obtained by merging the information stored in the knowledge base for all surface forms corresponding to all possible boundaries at the given location in text (in the above example, this is COMPOSITE("Australian", "Australian politics")).

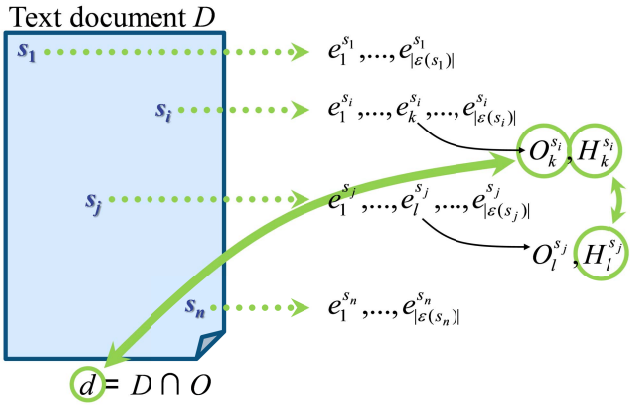

Figure 5. The disambiguation process computes the match between an entity candidate and the input document in the space of observable attributes, and the match between entity candidates of different surface forms in the space of latent attributes. In this figure, s_1 to s_n denote the surface forms extracted from the input text, e^s denote entities that a surface form s can be mapped to, while O^s and H^s denote the entity's observable and latent attributes, respectively.

2.3 Entity Disambiguation

The entity disambiguation stage employs all surface forms extracted in the text analysis stage in conjunction with the known mappings to entities and the attributes of those entities as stored in the knowledge base. As shown in Figure 5, the system employs the observables attributes to compute directly the similarity of an entity candidate with the input document (for example, the system computes how many of the known contexts of an entity appear in the input text), and the latent attributes to compute the matching between various hypothesized entities. Since computing the best pair-wise matching of latent attributes for all possible entity assignments is NP-complete, NEMO employs the technique described in [2] and [3] of building a document model by aggregating all attribute values for all possible entity disambiguations, and then computing the match between the latent attribute values of each entity candidate and this aggregated document model.

In total, the system uses 26 features computed as similarities in either the observable space or the latent space, which are further employed by a logistic-regression classifier to produce a score for each entity candidate. The candidate with the highest score in the entity mapping list of each surface form in the text is picked as the entity to be output by the system.

3. TRAINING BASED ON WIKIPEDIA

The system is trained by using a collection of paragraphs from Wikipedia, in which the interlinks created by the Wikipedia contributors are employed as labeled examples. A short sentence that contains one additional entity mention is added to each paragraph in the following manner: if the training paragraph originates from a Wikipedia page about an entity X then we add a sentence of the type ``This text is about [[X | *Entity_Form*(X)]]", in which *Entity_Form*(X) is an entity form known to have a possible mapping to the entity X (an actual example of training fragment is shown in Figure 6 for a fragment extracted from the Wikipedia page for the "Battle of Waterloo"). The entity form is chosen based on the distribution derived from Wikipedia over all surface forms of entity X.

During training, the system is allowed to extract its own surface forms, and output for those surfaces that match the original Wikipedia interlinks (and only for those) the feature values for all candidate entity disambiguations. We pair the correct disambiguation, as produced by the Wikipedia contributors through interlinking, with every of the other candidate entities produced by our disambiguation system and we train a logistic regression classifier by employing the feature value differences from each pair. In total, the training process for TAC 2013 and ERD 2014 made use of two million labeled pairs.

ORIGINAL TRAINING TEXT:

This text is about [[Battle of Waterloo|Waterloo]]. Allegedly, Napoleon tried to escape to North America, but the [[Royal Navy|Roya Navy]] was blockading French ports to forestall such a move. He finally surrendered to [[Captain (Royal Navy)|Captain]] [[Frederick Lewis Maitland (Royal Navy officer)|Frederick Maitland]] of [[Her Majesty's Ship|HMS]] ''[[HMS Bellerophon (1786)|Bellerophon]]'' on 15 July There was a campaign against French fortresses that still held out; [[Longwy|Longwy]] capitulated on 13 September 1815, the last to dc so. The [[Treaty of Paris (1815)|Treaty of Paris]] was signed on 20 November 1815. [[Louis XVIII of France|Louis XVIII]] was restored tc the throne of France, and Napoleon was exiled to [[Saint Helena|Saint Helena]], where he died in 1821.

THE INITIAL ANALYSIS OF THE TRAINING TEXT:

This text is about Waterloo. Allegedly, Napoleon tried to escape to North America, but the Royal Navy was blockading French ports tc forestall such a move. He finally surrendered to Captain Frederick Maitland of HMS ''Bellerophon'' on 15 July. There was a campaigr against French fortresses that still held out; Longwy capitulated on 13 September 1815, the last to do so. The Treaty of Paris was signec on 20 November 1815. Louis XVIII was restored to the throne of France, and Napoleon was exiled to Saint Helena, where he died in 1821.

Figure 6. Training example constructed from Wikipedia data, which contains a paragraph from the "Battle of Waterloo" article concatenated with an additional sentence that contains a reference to the entity discussed by the article. The original text above shows the Wikipedia markup for interlinks, as created by the Wikipedia contributors. Below is shown the analyzed version, in which the surface forms identified by the system are underlined. Those of them that match the anchors of the Wikipedia interlinks from the original text are employed as training data points.

```
Ocean car carriers Mitsui OSK Lines (MOL)
and Höegh Autoliners have decided to
consolidate their short sea shipping and
feeder operations in Europe with the
formation of a 50-50 joint venture, Euro
Marine Logistics (EML), which integrates
the existing European short sea and
logistics activities of both companies as
well as Euro Marine Carrier (EMC) and
Nissan Motor Car Carriers (NMCC), entities
in which the two companies are
shareholders.
```

Figure 7. Example of text from the development set, in which multiple companies together with corresponding acronyms are mentioned. Some of these (e.g. Nissan Motor Car Carriers) do not have entries in the target entity collection, but their mentions contain prefixes that could be mentions of other known entities (e.g. Nissan Motor).

4. ERD 2014 EVALUATION

4.1 System Adaptation to the ERD Guidelines

The reference collection of entities in the challenge is derived from the October 2013 Freebase/Wikipedia collections. Because we already had a version of the NEMO system, which participated in the September 2013 TAC evaluation, built and trained by using the Wikipedia dump from August 5, 2013, we employed this version for the ERD evaluation. The entities from the August 5, 2013 dump were mapped to the ERD reference collection by simply matching the titles from the two sets. Because of differences between the collections, this naïve process was not able to assign 28,036 of the Freebase IDs (roughly 1.2% of the IDs provided) to Wikipedia entities in NEMO's knowledge base.

The most challenging requirement of the ERD evaluation was to identify only the longest mentions of entities in the target collection, as NEMO was built to identify all mentions of entities, whether known or unknown. For example, in "Mitsubishi ASX3 1.8 DiD", the gold standard identifies "`Mitsubishi`" as the mention of the brand rather than the full string as the car model. Similarly, for the text "Nokia Maps", the gold standard contains the extraction "`Nokia`" instead of the full string, which gets mapped to "Here (Nokia)" by NEMO. Despite the fact that we obtained some overall performance improvements on the development set by tweaking the system to identify known substrings inside mentions of entities not in the target collection, we decided that the errors due to this tweak were too severe, as shown in the example in Figure 7. In the end, we applied the tweak for extracting the longest known surface of a target entity only for the cases in the substring-mentions refer to geo-political entities.

Additionally, to increase recall, we employed a more aggressive extraction of entities from capitalized contexts, and we also changed the handling of nationalities, with the regular version of the NEMO system discards. Instead, the ERD version of the system extracts the nationalities from text and handles them as composite surface forms that aggregate the information from the original surface form in text and most frequent entity form for the country corresponding to that nationality (for example, "Chinese" gets mapped to COMPOSITE("`Chinese`", "`China`")).

These changes resulted in a recall increase of more than 6 absolute points on the development set while keeping the precision at the same level, which translated in a gain in F-measure of 3 absolute points, as shown in Table 1.

Table 1. Performance of the NEMO system on the ERD development set before tuning it for the ERD task and after the changes made to follow the guidelines of the task.

ERD 2014 Dev. set	**Precision**	**Recall**	**F-measure**
Beginning performance	83.70	72.59	77.75
Final performance	83.33	78.92	81.07

4.2 Results

At the time of publishing these results, the annotated development (in its final version) and test sets had not been released yet due to the tight publication deadline immediately following the evaluation. Therefore, an in-depth analysis of the results is not available now, and will be published at a later time.

Table 2 shows the performance reported for the final version of the NEMO system on the test set, together with the reported ranges for each metric. The F-measure performance, which is the main metric in the ERD evaluation, as measured for both the development set and the test set, puts NEMO at the top of both development and test leaderboards. It is encouraging that the system obtains consistent precision on both (83.33% on development and 83.32% on test). The 10-point loss in recall on the test needs further investigation once the annotated data sets are released. We hypothesize that the system adaptation done based on the development set accounted for several annotation and linguistic phenomena that were specific to that particular data set. Because of the pooling strategy employed for evaluation, it is possible that the gold-standard annotation of the test contains types of entity extractions that did not surface in the development.

Table 2. Performance of the NEMO system on the ERD test set, together with the ranges reported for each metric, and the performance of the median system that entered the challenge.

ERD 2014 Test set	**Precision**	**Recall**	**F-measure**
NEMO	83.32	69.85	**75.99**
Median system	73.90	55.46	63.37
Range	51.94 – 87.56	30.35 – 71.16	44.56 – 75.99

5. CONCLUSION

The paper described the NEMO system for entity extraction and disambiguation, and its evaluation in the ERD 2014 challenge. The system obtained the best reported results for the long track on both the development and test sets.

6. REFERENCES

[1] Carmel, D., Chang, M.W., Gabrilovich, E., Hsu, P., and Wang, K. 2014. ERD 2014: Entity Recognition and Disambiguation Challenge. *SIGIR Forum 2014*, ACM.

[2] Cucerzan, S. 2007. Large Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP-CoNLL 2007*. 708–716.

[3] Cucerzan, S. 2012. The MSR System for Entity Linking at TAC 2012. *Text Analysis Conference 2012*. DOI= http://www.nist.gov/tac/publications/2012/participant.papers/MS_MLI.proceedings.pdf.

[4] Cucerzan, S. and Sil, A. 2013. The MSR Systems for Entity Linking and Temporal Slot Filling at TAC 2013. *Text Analysis Conference 2013*. DOI= http://www.nist.gov/tac/publications/2013/participant.papers/MS_MLI.proceedings.pdf.

[5] Mayfield, J., Artiles, J., and Dang, H. T. 2012. Overview of the TAC2012 Knowledge Base Population. *Text Analysis Conference 2012*. http://www.nist.gov/tac/publications/2012/additional.papers/KBP2012_overview.notebook.pdf