Named Entity Disambiguation for Noisy Text

YOTAM ESHEL, Technion
NOAM COHEN, Technion
KIRA RADINSKY, eBay
SHAUL MARKOVITCH, Technion
IKUDA YAMADA, Studio Ousia
OMER LEVI, University of Washington

Named Entity Disambiguation (NED) is the task of linking mentions of entities in text to a given knowledge base, such as Freebase or Wikipedia. NED is a key component in Entity Linking (EL) systems, focusing on the disambiguation task itself, independently from the tasks of Named Entity Recognition (detecting mention bounds) and Candidate Generation (retrieving the set of potential candidate entities). NED has been recognized as an important component in NLP tasks such as semantic parsing [2].

Current research on NED is mostly driven by a number of standard datasets, such as CoNLL-YAGO [6], TAC KBP [7] and ACE [1]. These datasets are based on news corpora and Wikipedia, which are naturally coherent, well-structured, and rich in context. Global disambiguation models [4, 5, 9] leverage this coherency by jointly disambiguating all the mentions in a single document. However, domains such as web-page fragments, social media, or search queries, are often short, noisy, and less coherent; such domains lack the necessary contextual information for global methods to pay off, and present a more challenging setting in general.

In this work, we investigate the task of NED in a setting where only *local* and *noisy* context is available. In particular, we create a dataset of 3.2M short text fragments extracted from web pages, each containing a mention of a named entity. Our dataset is far larger than previously collected datasets, and contains 18K unique mentions linking to over 100K unique entities. We have empirically found it to be noisier and more challenging than existing datasets. For example:

> "I had no choice but to experiment with other indoor games. I was born in Atlantic City so the obvious next choice was **Monopoly**. I played until I became a successful Captain of Industry."

This short fragment is considerably less structured and with a more personal tone than a typical news article. It references the entity *Monopoly_(Game)*, however expressions such as "experiment" and "Industry" can distract a naive disambiguation model because they are also related the much more common entity *Monopoly* (economics term). Some sense of local semantics must be considered in order to separate the useful signals (e.g. "indoor games", "played") from the noisy ones.

We therefore propose a new model that leverages local contextual information to disambiguate entities. Our neural approach (based on RNNs with attention) leverages the vast amount of training data in WikilinksNED to learn representations for entity and context, allowing it to extract signals from noisy and unexpected context patterns.

While convolutional neural networks [3, 10] and probabilistic attention [8] have been applied to the task, this is the first model to use RNNs and a neural attention model for NED. RNNs account for the sequential nature of textual context while the attention model is applied to reduce the impact of noise in the text.

XXXX-XXXX/2017/4-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Our experiments show that our model significantly outperforms existing state-of-the-art NED algorithms on WikilinksNED, suggesting that RNNs with attention are able to model short and noisy context better than current approaches. In addition, we evaluate our algorithm on CoNLL-YAGO [6], a dataset of annotated news articles. We use a simple domain adaptation technique since CoNLL-YAGO lacks a large enough training set for our model, and achieve comparable results to other state-of-the-art methods. These experiments highlight the difference between the two datasets, indicating that our NED benchmark is substantially more challenging.

Additional Key Words and Phrases: Entity Linking, Entity Disambiguation

ACM Reference format:
Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuda Yamada, and Omer Levi. 2017. Named Entity Disambiguation for Noisy Text. 1, 1, Article 1 (April 2017), 2 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

REFERENCES

[1] Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. 2010. *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources.* Coling 2010 Organizing Committee, Chapter Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia, 19–27. http://aclweb.org/anthology/W10-3503

[2] Jonathan Berant and Percy Liang. 2014. Semantic Parsing via Paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 1415–1425. https://doi.org/10.3115/v1/P14-1133

[3] Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 1256–1261. https://doi.org/10.18653/v1/N16-1150

[4] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. 2016. Collective Entity Resolution with Multi-Focal Attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 621–631. https://doi.org/10.18653/v1/P16-1059

[5] Zhaochen Guo and Denilson Barbosa. 2014. Entity Linking with a Unified Semantic Representation. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion).* ACM, New York, NY, USA, 1305–1310. https://doi.org/10.1145/2567948.2579705

[6] Johannes Hoffart, Amir Mohamed Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 782–792. http://aclweb.org/anthology/D11-1072

[7] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, Vol. 3. 3–3.

[8] Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A Selective Context Model for Entity Resolution. *Transactions of the Association of Computational Linguistics* 3 (2015), 503–515. http://aclweb.org/anthology/Q15-1036

[9] Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized Page Rank for Named Entity Disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 238–243. https://doi.org/10.3115/v1/N15-1026

[10] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015.* 1333–1339. http://ijcai.org/Abstract/15/192