Human-In-The-Loop Entity Linking for Low Resource Domains

Jan-Christoph Klie **Richard Eckart de Castilho** **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technical University of Darmstadt, Germany
www.ukp.tu-darmstadt.de

Abstract

Entity linking (EL) is concerned with disambiguating entity mentions in a text against a knowledge base (KB). To quickly annotate texts with EL in low-resource domains and noisy text, we present a novel Human-In-The-Loop EL approach. We show that it greatly outperforms a strong baseline in simulation. In a user study, annotation time is reduced by 35 % compared to annotating without interactive support; users report that they strongly prefer our new approach. An open-source and ready-to-use implementation based on the text annotation platform INCEpTION[1] is made available [2].

1 Introduction

Entity linking (EL) describes the task of disambiguating entity mentions in a text by linking them to a knowledge base (KB), e.g. the text span *Earl of Orrery* can be linked to the KB entry *John Boyle, 5^{th} Earl of Cork*, thereby disambiguating it. EL is highly relevant in many fields like digital humanities, classics, technical writing or biomedical sciences for applications like search (Meij et al., 2014), semantic enrichment (Schlögl and Lejtovicz, 2017) or information extraction (Nooralahzadeh and Øvrelid, 2018).

In these scenarios, the first crucial step is typically to annotate data. Manual annotation is laborious and often prohibitively expensive. To improve annotation speed and quality, we have developed a novel Human-In-The-Loop (HITL) entity linking approach. It helps annotators finding entity mentions in the text and linking them to the correct knowledge base entries. The more mentions get linked over time, the better the annotation support will be.

We demonstrate the effectiveness of our approach with extensive simulation as well as a user study on different, challenging datasets. We have implemented our approach based on the open-source annotation platform INCEpTION (Klie et al., 2018) and publish all datasets and code.

2 Implementation

Entity linking describes the task of disambiguating mentions in a text against a knowledge base. Manual annotation of EL consists of three steps (Shen et al., 2015). First, the annotator selects a span that contains an entity. Then, they search for the correct entity in a KB. These search results are reranked to rank more suitable candidates higher. Each candidate from the knowledge base is assumed to have a label and a description.

To speed up this annotation process, we support users twofold: To find suitable spans, we provide *recommenders* that suggest potential entity spans. They can also classify these entity spans (e.g. as person, location, etc.). These recommenders learn from new annotations and are retrained in the background. For candidate ranking, we follow Zheng et al. (2010) and model it as a learning-to-rank problem: given a marked span, search query and a list of candidates, sort the candidates so that the most relevant candidate is at the top. By selecting an entity label from the candidate list, users express that the selected one was preferred over all other candidates. These preferences are used to train state-of-the-art pairwise learning-to-rank models from the literature: the gradient boosted trees variant `LightGBM` (Ke et al., 2017) and `RankSVM` (Joachims, 2002). The continuously updated models improve over time with an increasing number of annotations. As input features, we use different similarity measures between the marked span and the candidate label, between the spans' context and the candidate description as well as dense word and sentence embeddings of the descriptions.

[1] https://inception-project.github.io
[2] https://github.com/UKPLab/acl2020-interactive-entity-linking

Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances, pages 41–43

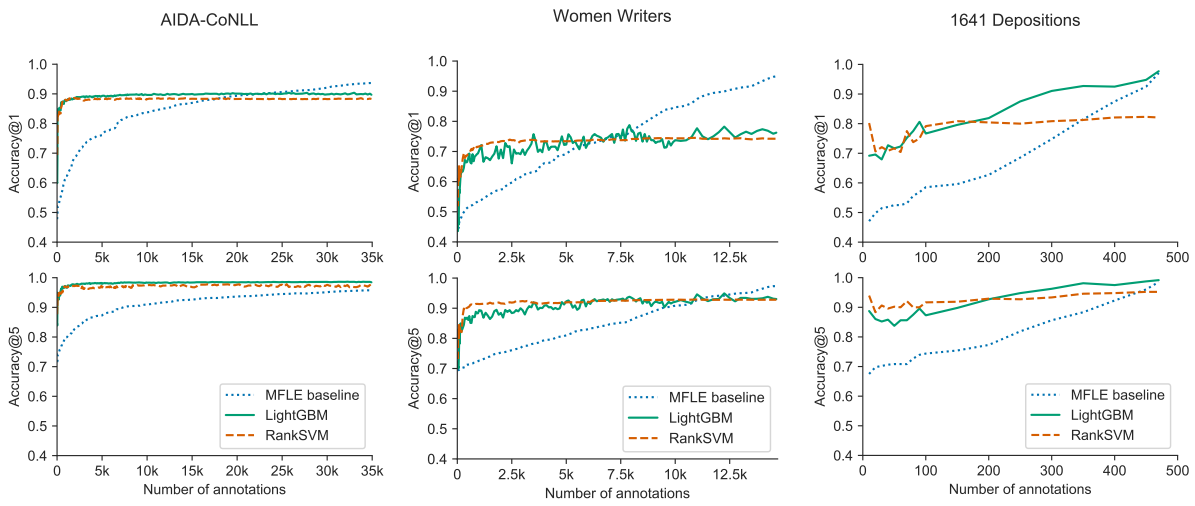

Figure 1: Human-in-the-loop simulation results for our three datasets and models. One can see that the model achieves good Accuracy@5 with only a few annotations, especially for the `RankSVM`.

Datasets We use the following three datasets for validating our approach: 1) the `AIDA-YAGO` state-of-the art dataset introduced by Hoffart et al. (2011). 2) `Women Writers Online`[3] is a collection of texts by pre-Victorian women writers. It includes texts on a wide range of topics and from various genres including poems, plays, and novels. 3) The `1641 Depositions`[4] contain legal texts in form of court witness statements recorded after the Irish Rebellion of 1641.

3 Experiments

To validate our approach, we simulate a user annotating with our HITL ranker. Then, we conduct a user study to test it in a real-life setting. Similar to other work on EL, our main metric for ranking is accuracy. We also measure Accuracy@5, as our experiments showed that users can quickly scan and select the right entity from a list of five elements.

Simulation Fig. 1 depicts the simulation results. All models outperform a majority baseline over most of the annotation process. It can be seen that both of our used models achieve high performance even if trained on very few annotations. The `RankSVM` handles low data better than `LightGBM`, but quickly reaches its peak performance due to it being a linear model. This potentially allows to first use a `RankSVM` for the cold start and when enough annotations are made, `LightGBM`, thereby combining the best of both.

User Study In order to validate the viability of our approach in a realistic scenario, we conduct a user study. For that, we augmented the already existing annotation tool INCEpTION (Klie et al., 2018) with our Human-In-The-Loop entity ranking and automatic suggestions. We let five users re-annotate parts of the `1641` corpus. We compare two configurations: one uses our reranking, one uses the default ranking. We randomly selected eight documents which we split in two sets of four documents. We measure annotation time, number of suggestions used and search queries performed. The evaluation of the user study shows that using our approach, users on average annotated 35% faster and needed 15% fewer search queries.

4 Conclusion

We presented a domain-agnostic annotation approach for annotating entity linking for low-resource domains. It consists of two main components: recommenders that are algorithms that suggest potential annotations to users and a ranker that, given a mention span, ranks potential entity candidates so that they show up higher in the candidate list, making it easier to find for users. Both systems are retrained whenever new annotations are made, forming the Human-In-The-Loop. In a user study, results show that users prefer our approach compared to the typical annotation process; annotation speed improves by around 35% when using our system relative to using no reranking support.

[3] https://www.wwp.northeastern.edu/wwo
[4] http://1641.tcd.ie/

References

Johannes Hoffart, Mohamed Amir 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 EMNLP'11*, pages 782–792.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, pages 133–142.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Edgar Meij, Krisztian Balog, and Daan Odijk. 2014. Entity linking and retrieval for semantic search. In *Proceedings of the 7th ACM international conference on Web search and data mining - WSDM '14*, pages 683–684.

Farhad Nooralahzadeh and Lilja Øvrelid. 2018. SIRIUS-LTG: An Entity Linking Approach to Fact Extraction and Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 119–123.

Matthias Schlögl and Katalin Lejtovicz. 2017. APIS - Austrian Prosopographical Information System. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Learning to Link Entities with Knowledge Base. In *Prooceedings of NAACL-HLT'10*, pages 483–491.