Riches of the Poor: Using Crummy Entity Linkers for Interactive Search in Digital Humanities

Olieman, A.M.; Beelen, K.; van Lange, M.; Kamps, J.; Marx, M.J.

Riches of the Poor: Using Crummy Entity Linkers for Interactive Search in Digital Humanities *

Alex Olieman[1,2] Kaspar Beelen[1] Milan van Lange[3] Jaap Kamps[1] Maarten Marx[1]

[1] University of Amsterdam, The Netherlands, {olieman,k.beelen,kamps,maartenmarx}@uva.nl
[2] Stamkracht BV, Amsterdam, The Netherlands, alex@stamkracht.com
[3] NIOD Institute for War, Holocaust, and Genocide Studies, Amsterdam, The Netherlands, m.van.lange@niod.knaw.nl

1 INTRODUCTION

The traditional approach for evaluating Entity Linking (EL) systems employs metrics borrowed from Information Retrieval, *precision* and *recall*, to summarize system performance and to aid in their comparison. Consequently, the notion of improving EL technology is closely tied to demonstrating an increased performance in terms of precision and recall. Recent evaluation efforts, utilizing multiple benchmark datasets, have however shown that a good performance on one dataset often does not generalize to others [2]. This limitation of current EL benchmarks, we argue, limits the usefulness of performance statistics for application developers who would like to include EL in their software.

2 ENTITY LINKING EVALUATION

There is a poor understanding of the state-of-the-art in Entity Linking [2, 4], which can be partly blamed on the desire for generality in the task definition and its associated objects. How might we then go about demonstrating how effective or useful EL technology can be, in spite of the issues with our metrics and benchmarks? It would be beneficial, in terms of impact and continued funding, to show that EL technology can already be valuable for end-users in its current form.

3 SEARCHING FOR ENTITIES BY THE DOZEN

In this paper [3] we introduced WideNet, an entity-based interactive search tool, specifically tailored to assist historians with the exploration of large document collections. Users query the tool with a category of entities and the UI subsequently shows, per subcategory, which entities are mentioned in the corpus, and how frequently, as well as which entities did not occur (see Figure 1). It also displays a list of preview results, showing limited context, to offer quick clues about the relevance of the category.

4 CONCLUSION

The main motivation behind the creation of WideNet was to show that *(a)* even imperfect technologies such as the state-of-the-art Entity Linking systems can support useful niche applications, and *(b)* tool design should take into account the methodological practices of the scholar. We demonstrated in [3] how the interface supports historians by providing them with a holistic overview of references to complex phenomena such as historical events. Moreover, the grouped presentation of search results with different levels of context dampens frequency in favor of variety: long tail entities become more visible while highly frequent but irrelevant results can easily be discarded.

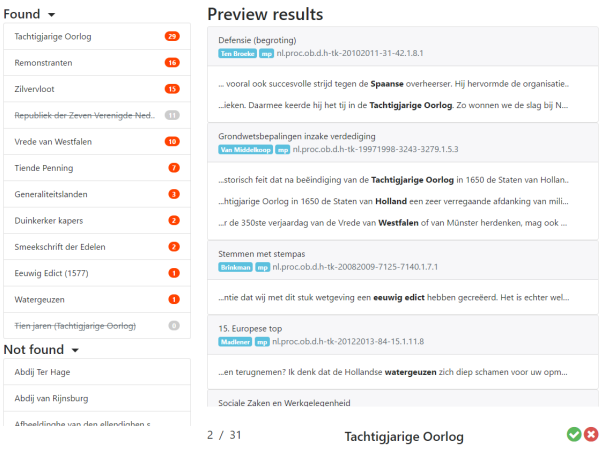

Figure 1: Assessing the relevance of categories and entities.

Our general conclusion is that the use case and application crucially determine whether the quality of an EL system is good enough or not: the same EL system may be useless for one task but very useful for another. Hence, paraphrasing a famous paper on early machine translation [1]: there are good applications of crummy entity linkers.

Acknowledgments. This research was supported in part by Netherlands Organization for Scientific Research (ExPoSe, NWO CI # 314.99.108; DiLiPaD, NWO Digging into Data # 600.006.014).

REFERENCES

[1] Kenneth W. Church and Eduard H. Hovy. 1993. Good Applications for Crummy Machine Translation. *Machine Translation* 8 (1993), 239–258.
[2] Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design Challenges for Entity Linking. *Transactions of ACL* 3, 2011 (2015), 315–328.
[3] Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good Applications for Crummy Entity Linkers? The Case of Corpus Selection in Digital Humanities. In *Proceedings of SEMANTiCS 2017.* https://doi.org/10.1145/3132218.3132237 arXiv:1708.01162
[4] Marieke van Erp, Pablo N. Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, and Giuseppe Rizzo. 2016. Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. In *Proceedings of LREC '16.* 4373–4379.

*This is an extended abstract of [3]

DIR'17, November 2017, Hilversum, Netherlands