KG-Agnostic Entity Linking Orchestration

Kristian Noullet[1][0000−0002−4916−9443]

Karlsruhe Institute of Technology, Kaiserstr. 89, 76133 Karlsruhe, Germany
kristian.noullet@kit.edu

Abstract. The domain of Entity Linking (EL) has been researched thoroughly, resulting in numerous approaches and a level of maturity in the field. In this dissertation, we aim at (1) creating an orchestrated, continuously-improving system through addition of approaches both existing and ones that are yet to come. On the same note, our framework shall (2) allow for Knowledge Graph (KG)- and potentially language-agnostic EL processes through cross-KG techniques, enabling further exploration of KG-dependence in the domain of EL. Finally, we will boost the ease of (3) reproducibility and (4) comparison between EL systems, their underlying techniques, as well as orchestration frameworks as a whole.

Keywords: NLP · Entity Linking · Orchestration · KG-Agnosticity.

1 Problem Statement

In the world of EL, there exist a plethora of approaches attempting to solve the issue of detecting and correctly disambiguating *mentions* from plain text documents. Many of these approaches are - at least primarily - bound to specific KGs, making system comparison problematic and portability of results to other knowledge bases questionable. Additionally, - with a few exceptions - utilizing only specific parts of a given system's pipeline to further be processed by other systems is relatively hard, consequently increasing a researcher's load even when intending to only research specific steps. Thus, proper system comparability and *step*[1]-related result explanation become potentially tedious tasks, for instance when attempting to solely improve Entity Disambiguation (ED), whilst intending to keep existing Mention Detection (MD) and Candidate Generation (CG) techniques constant.

Inter-System vs. Intra-System We broadly differentiate between two types of systems utilised for our orchestration: inter- and intra-system. The former refers to orchestration across multiple full-fledged EL systems, applying End-to-End EL systems' results and combining them in specific ways. Among these are approaches, such as Babelfy [15], DBpedia Spotlight [14], DoSeR [26] and MAG [16] - to name a few. These frameworks loosely follow the steps of MD, CG and ED,

[1] Referring to singular steps of an EL pipeline, most commonly: Mention Detection (MD), Candidate Generation (CG) or Entity Disambiguation (ED)

meaning they may be split into modules for use in intra-system orchestration. In contrast to the inter-system approach, intra-system orchestration refers to the application of sub-system components interacting in order to improve intermediary results, potentially yielding an overall qualitatively further developed final output. The advantage of accessing specific processing steps prior to final result output lies in allowing for further dynamic access and potentially utilizing highly-developed techniques without having to re-compute nor re-develop certain approaches. As intermediary steps are improved upon, result performance shall ultimately increase as well. Among others, with the rise of neural EL approaches, not every EL system may be regarded as a succession of steps of MD, CG and ED. For instance, Broscheit [3] and Kolitsas et al. [13] have developed joint-EL techniques, making splitting these into encapsulated stand-alone subsystems difficult.

As such, we define the following problems our dissertation intends to alleviate and potentially resolve:

P1 Multitude of 'Closed' Heterogeneous Systems.

P2 Lack of Cross-KG and Robust KG-Agnostic EL Orchestration.

P3 EL System Evaluation (Comparability, Result Reproducibility and Explainability).

1.1 Multitude of 'Closed' Heterogeneous Systems.

While there exist a multitude of systems to choose from - each with their own set of peculiarities -, to the best of our knowledge, a centralised way of utilising existing approaches' underlying steps without considerable development effort has yet to be released. In the case of open source projects, development and adjustments may yield wanted results, but the accompanying effort and required underlying data structures render the task tedious, if not infeasible. Further, taking into account licensing as well as privacy concerns - prohibiting distribution - regarding potentially-utilised knowledge sources, makes it apparent where limits of system, as well as result reproducibility may be reached. Therefore, our goal is to create an orchestration framework, allowing for ease of research and development in the domain of EL without having to worry about every step. Alike GERBIL [19], we intend to centralise and simplify the use of existing approaches, while our goal is majorly contrasted by the aim of using concurrent systems to a higher level of granularity, not for evaluation, but for combined inter- as well as intra-system orchestration. Consequently, as a major byproduct of this PhD. proposal, we will provide interfaces for ease of implementation, a REST API-based endpoint to execute existing (and register novel) approaches, additionally to supplying a training & testing platform in order to boost existing EL research, as well as to apply various techniques for a holistic orchestration approach.

1.2 Lack of Cross-KG and Robust KG-Agnostic EL Orchestration.

Although there exist KG-Agnostic EL frameworks, to the best of our knowledge existing orchestration approaches do not approach the domain of working with multiple KGs simultaneously (Cross-KG) nor work for potentially "any" wanted knowledge base (KG-Agnostic). The lack of system portability between KGs can be problematic for evaluation, maintenance as also ensuring the use of a system in the long run - either due to lack of community interest for a given KG or changes in regards to licensing. In the past, EL approaches have strongly relied on Wikipedia/DBpedia [19], but more recently, research has shifted slightly in favor of developing systems for Wikidata [8, 21, 6, 17], among others. Initially, we will be working with the baseline of translating systems' annotations to a common basis through `owl:sameAs` predicate links, allowing for orchestration through supervised learning approaches with help of gold standards. Further down the line, our intention is to have similar entity detection across KGs through entity alignment techniques [25, 23].

1.3 EL System Evaluation.

GERBIL [19] has greatly contributed to facilitating evaluation of EL systems, as well as comparing their result metrics to one another. Unfortunately, it does not allow for checking results directly - unless an endpoint is set up as a relay in a man-in-the-middle type of approach -, but it serves well its purpose of properly allowing overall performance comparison between systems. Spiritually, we follow a similar route for **system evaluation**, but rather go on a more granular level of the common EL pipeline for the purpose of combining a multitude of (sub-)approaches and learning optimal orchestration settings and potential configurations for specific domains. Unfortunately, to the best of our knowledge, existing approaches do not allow for the ease of **result reproducibility** to take place for MD, CG and ED separately. For the purpose of boosting ease of research focusing solely on specific parts of existing pipelines, we wish to remediate this lack with our system by defining common protocols for each of these steps. Additionally, we shall do so for their combinations as well, in order to further allow highly configurable variations in future systems to arise. Allowing for common entity linking pipelines' modularization further allows for easier **result explainability** by exploring output variations due to sub-system mechanism swaps. Further, thanks to our approach aiming at being KG-Agnostic, we shall facilitate domain-dependent explainability greatly.

2 Research Questions & Hypotheses

RQ1 How can we leverage multiple KG systems' advantages for various settings?
RQ2 When does which system(s) yield optimal results?
RQ3 To what extent do domains impact EL results?
RQ4 How to utilise KG-bound EL systems for agnostic tasks?

RQ5 How can we achieve and evaluate the level of KG-Agnosticity of various systems?

Consequent to our research questions, we make the following hypotheses:

H1 Different linking systems work to different extents based on target and underlying KGs.
H2 Domains of plain texts affect the quality of results strongly (e.g. political vs. geographical, see Obama[2] and Obama[3]).
H3 Constraining a given system to a specific KG potentially lowers quality of results.
H4 EL research is dampened by a development overhead, rendering system comparability and reproducibility difficult, especially when varying between underlying knowledge bases.

3 Related Work

As can be seen in GERBIL[19, 2] and [22], EL is a mature area of research, including a multitude of approaches. These are based on various techniques. We broadly differentiate between graph-based and machine learning-based EL frameworks. On one hand, you have graph-based approaches, such as Babelfy [15] and MAG [16]. These make use of graph algorithms to find contextually similar entities with the assumption that (candidate) entities which are located in close proximity to each other within a KG, increases the likelihood of appearing together within texts. As such, Babelfy applies a densest subgraph algorithm, whereas MAG employs a breadth-first search style of search for entities within a defined KG. Other types of EL approaches include machine learning-based and neural techniques. Among their ranks may be listed [3, 5, 10, 13, 24, 26]. In [3], Broscheit utilizes some of the explosively-popular BERT [9] models to achieve neural end-to-end EL, ideologically similarly to Kolitsas et al.'s [13] joint neural EL approach. Main contributions relating to our area of EL system orchestration were brought by [4, 7, 20, 12]. In [12], João et al. present an end-to-end supervised learning approach to orchestration on 3 systems (Babelfy [15], TagMe [11] and Ambiverse [1]), analyzing performance for different settings, such as binary and multi-label classification. In contrast, Canale et al's. [4] developed system applies a corpus-learnt voting scheme to annotators, while Corcoglioniti et al. employ a more constant majority-based voting approach with MicroNeel [7]. The former learn weights for their voting application, whereas the latter predominantly focuses on short documents, resolving potential conflicts with predetermined priorities. Finally, we consider GERBIL [19] to be a spiritual predecessor to a certain extent due to (1) its goal of greatly simplifying processes relating to EL; (2) providing interfaces and code templates for ease of implementation; (3) aggregating a multitude of approaches; and (4) unifying approaches (for evaluation), therewith improving system comparison.

[2] https://en.wikipedia.org/wiki/Obama,_Fukui
[3] https://en.wikipedia.org/wiki/Barack_Obama

4 Approach

In order to achieve a sensible and highly extensible orchestration system, we have started by implementing our own KG-Agnostic embeddings-based EL approach, codenamed Agnos. Through its use of multiple KGs and various mechanisms for MD, CG and ED, we identified the degree of generalization, data structures and processes must provide in order to allow for maximal inter-compatibility with existing systems, whilst simultaneously minimizing the loss of linker-specific information. Both inter-system and intra-system compatibility is warranted and an initial REST API is provided. The code base for our EL approach is accessible at `https://git.scc.kit.edu/wf7467/agnos_mini`, additionally to our project including initial orchestration capabilities at `https://github.com/kmdn/Agnos_mini`.
Each step from mention detection to pruning is handled by individual interfaces, allowing for definition of REST API-based connectivity for modularity of each desired step.
Once we have a relatively mature system set up, we will initially apply supervised learning techniques for our cross-KG approach. These will be supported by `owl:sameAs` predicates, allowing for a transparent translation layer of interconnections between KGs. Further down the line, we will additionally shift to exploiting KG alignment techniques to further facilitate the use of a plethora of KGs - both commonly available, as well as custom and potentially private ones. Due to our learning approach, we will further subdivide our experiments into various domains in order to properly assess their relevance and influence on result quality.

5 Evaluation

We shall assess our work from an EL point of view by using GERBIL along with its performance metrics for evaluation - if data sets fulfill required parameters, such as that of size and domain relevance for our experiments. Unfortunately, GERBIL requires a reference knowledge base for data sets which may increase difficulty for a KG-Agnostic evaluation. Therefore, we will either develop other data sets or use existing ones (e.g. [18]) for cross-KG evaluation, depending on whether appropriate ones may be located. Further, we will proceed with testing our orchestration system while taking into consideration its own orchestration with other end-to-end frameworks to explore whether major improvements may be achieved on an even more *meta*-level. Whilst our primary objective is to further boost EL result quality, it is important to not only evaluate our system as an EL framework, but also as a platform. Therefore, we will conduct user studies to evaluate useability as well as potential time gain achieved through our system, along the lines of `https://github.com/dice-group/gerbil/blob/master/documentation/survey.csv`.
Finally, we will attempt to reach more detailed conclusions in terms of different frameworks' domain dependence as well as analyse areas of both opportunity and potential system limitations due to our enabled enhanced system granularity.

6 Future Work

In terms of further research to come, we will develop our framework for purposes of maximising intercompatibility of EL systems through usage of the NLP Interchange Format (NIF[4]). Additionally to applying supervised classification techniques, we will attempt to additionally remodel the problem in order to tackle it from different angles, perchance enabling the use of large corpora without the necessity of tedious creation of silver and gold standards. We will further refine our framework, publishing the finalised interfaces and code templates, once we consider them mature enough. Alongside framework developments, we will set up an openly-available REST API, along with a variety of configurations to choose from for execution of specific EL systems, as well as orchestration baselines.
For evaluation purposes, we would like to increase the range of data sets applied to our task. We will be targetting various gold standards of sufficient size (e.g. CoNLL2011), but would like to contribute to the community by creating our own, as well as separating these into various domains to enable the impact of domain to the field of research.
On a similar note, we will be training our system - among others - on specific domains and evaluating them on the source domain, as well as various target domains in an attempt to evaluate domain dependency.
Finally, as our project progresses, we will switch from relatively simple `owl:sameAs` link usage to additionally making use of alignment techniques, increasing system agnosticity and likely yielding advantages from a multitude of approaches, aggregated into a single framework.

References

1. AmbiverseNLU: A Natural Language Understanding suite by Max Planck Institute for Informatics. https://www.mpi-inf.mpg.de/ambiverse-nlu/, accessed: 2020-05-30
2. GERBIL Experiment Configuration. http://gerbil.aksw.org/gerbil/config, accessed: 2020-05-30
3. Broscheit, S.: Investigating entity knowledge in BERT with simple neural end-to-end entity linking. CoRR **abs/2003.05473** (2020), https://arxiv.org/abs/2003.05473
4. Canale, L., Lisena, P., Troncy, R.: A novel ensemble method for named entity recognition and disambiguation based on neural network. In: International Semantic Web Conference. pp. 91–107. Springer (2018)
5. Cao, Y., Hou, L., Li, J., Liu, Z.: Neural collective entity linking. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. pp. 675–686. Association for Computational Linguistics (2018), https://www.aclweb.org/anthology/C18-1057/

[4] http://aksw.org/Projects/NIF.html

6. Cetoli, A., Bragaglia, S., O'Harney, A.D., Sloan, M., Akbari, M.: A neural approach to entity linking on wikidata. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11438, pp. 78–86. Springer (2019). https://doi.org/10.1007/978-3-030-15719-7_10, https://doi.org/10.1007/978-3-030-15719-7_10
7. Corcoglioniti, F., Aprosio, A.P., Nechaev, Y., Giuliano, C.: MicroNeel: Combining NLP Tools to Perform Named Entity Detection and Linking on Microposts. In: Basile, P., Corazza, A., Cutugno, F., Montemagni, S., Nissim, M., Patti, V., Semeraro, G., Sprugnoli, R. (eds.) Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016. CEUR Workshop Proceedings, vol. 1749. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1749/paper_010.pdf
8. Delpeuch, A.: Opentapioca: Lightweight entity linking for wikidata. CoRR **abs/1904.09131** (2019), http://arxiv.org/abs/1904.09131
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Fang, Z., Cao, Y., Zhang, D., Li, Q., Zhang, Z., Liu, Y.: Joint entity linking with deep reinforcement learning. CoRR **abs/1902.00330** (2019), http://arxiv.org/abs/1902.00330
11. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In: Huang, J., Koudas, N., Jones, G.J.F., Wu, X., Collins-Thompson, K., An, A. (eds.) Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010. pp. 1625–1628. ACM (2010). https://doi.org/10.1145/1871437.1871689
12. João, R.S., Fafalios, P., Dietze, S.: Better together: an ensemble learner for combining the results of ready-made entity linking systems. In: Hung, C., Cerný, T., Shin, D., Bechini, A. (eds.) SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 - April 3, 2020. pp. 851–858. ACM (2020). https://doi.org/10.1145/3341105.3373883
13. Kolitsas, N., Ganea, O., Hofmann, T.: End-to-end neural entity linking. In: Korhonen, A., Titov, I. (eds.) Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018. pp. 519–529. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/k18-1050
14. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Ghidini, C., Ngomo, A.N., Lindstaedt, S.N., Pellegrini, T. (eds.) Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011. pp. 1–8. ACM International Conference Proceeding Series, ACM (2011). https://doi.org/10.1145/2063518.2063519
15. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics **2**, 231–244 (2014)
16. Moussallem, D., Usbeck, R., Röder, M., Ngomo, A.N.: MAG: A multilingual, knowledge-base agnostic and deterministic entity linking approach. In: Corcho,

Ó., Janowicz, K., Rizzo, G., Tiddi, I., Garijo, D. (eds.) Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017. pp. 9:1–9:8. ACM (2017). https://doi.org/10.1145/3148011.3148024

17. Mulang, I.O., Singh, K., Vyas, A., Shekarpour, S., Sakor, A., Vidal, M., Auer, S., Lehmann, J.: Context-aware entity linking with attentive neural networks on wikidata knowledge graph. CoRR **abs/1912.06214** (2019), http://arxiv.org/abs/1912.06214
18. Noullet, K., Mix, R., Färber, M.: KORE 50^{dywc}: An evaluation data set for entity linking based on dbpedia, yago, wikidata, and crunchbase. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020. pp. 2389–2395. European Language Resources Association (2020), https://www.aclweb.org/anthology/2020.lrec-1.291/
19. Röder, M., Usbeck, R., Ngomo, A.N.: GERBIL - benchmarking named entity recognition and linking consistently. Semantic Web **9**(5), 605–625 (2018). https://doi.org/10.3233/SW-170286
20. Ruiz, P., Poibeau, T.: Combining open source annotators for entity linking through weighted voting. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 211–215 (2015)
21. Sakor, A., Singh, K., Patel, A., Vidal, M.: FALCON 2.0: An entity and relation linking tool over wikidata. CoRR **abs/1912.11270** (2019), http://arxiv.org/abs/1912.11270
22. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028
23. Sun, Z., Hu, W., Zhang, Q., Qu, Y.: Bootstrapping entity alignment with knowledge graph embedding. In: IJCAI. pp. 4396–4402 (2018)
24. Xue, M., Cai, W., Su, J., Song, L., Ge, Y., Liu, Y., Wang, B.: Neural collective entity linking based on recurrent random walk network learning. In: Kraus, S. (ed.) Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. pp. 5327–5333. ijcai.org (2019). https://doi.org/10.24963/ijcai.2019/740
25. Zhu, H., Xie, R., Liu, Z., Sun, M.: Iterative entity alignment via joint knowledge embeddings. In: IJCAI. pp. 4258–4264 (2017)
26. Zwicklbauer, S., Seifert, C., Granitzer, M.: Doser - A knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9678, pp. 182–198. Springer (2016). https://doi.org/10.1007/978-3-319-34129-3_12, https://doi.org/10.1007/978-3-319-34129-3_12