CLiT: Combining Linking Techniques for Everyone

Kristian Noullet[0000−0002−4916−9443], Samuel Printz[0000−0003−1336−8536], and Michael Färber[0000−0001−5458−8645]

Karlsruhe Institute of Technology (KIT)
{kristian.noullet,michael.faerber}@kit.edu
samuel.printz@student.kit.edu

Abstract. While the path in the field of Entity Linking (EL) has been long and brought forth a plethora of approaches over the years, many of these are exceedingly difficult to execute for purposes of detailed analysis. In many cases, implementations are available, but far from being a *plug-and-play* experience. We present Combining Linking Techniques (CLiT), a framework with the purpose of executing singular linking techniques and complex combinations thereof, with a higher degree of reusability, reproducibility and comparability of existing systems in mind. Furthermore, we introduce protocols for the exchange of sub-pipeline-level information with existing and novel systems for heightened out-of-the-box compatibility. Among others, our framework may be used to consolidate multiple systems in combination with meta learning approaches and increase support for backwards compatibility of existing benchmark annotation systems.

Keywords: Entity Linking · Meta-Learning · Reproducibility · NLP · Semantic Web.

1 Introduction

The domain of Entity Linking (EL) deals with the interlinkage of textual mentions in text-based documents to corresponding entities in knowledge graphs. Researching and developing EL systems is a highly time-consuming process, encompassing a multitude of considerations at each step, including a plethora of moving parts – each capable of affecting the final results. Therefore, singling out the reason for the success – or failure – through ablation studies oftentimes constitutes a complex task, as any part of the processing pipeline may entail major changes. Consequently, comparability to other systems is effectively rendered *impossible* without tremendous research efforts. Even if such efforts are put in for a single system, being able to make use of these for novel research may pose an issue. In order to address these issues, we have worked on developing CLiT as a highly modular and flexible framework, allowing for an ease of adoption into existing systems and ones to come.

While research efforts allowing for performance evaluation of annotation tools have been developed, easing the centralised execution of said systems for the

purpose of further processing results has been mostly untouched. We intend to further extend the philosophy of increasing comparability between annotators through predefined evaluation data sets and computed metrics presented in [10,7,8], by enabling the use of complex workflows and in-depth analyses. In contrast to GERBIL [10], CLiT executes pipelines to a more granular degree, as well as combine the aforementioned, creating novel workflows. Further, the Silk framework [11] provides a *comparable* workflow, made up of various kinds of components interacting with each other. While the level of granularity is similar, to the best of our knowledge, CLiT introduces a larger degree of distributed customization and specialises more directly on the task of entity linking. In alignment with the vision of the Web of Data, all of our workflow's components and output provide and consume machine-readable data formats, in particular NIF 2.0 and JSON.

We intend to lead the research towards being able to answer the following research questions:

1. How can the research community **leverage** (sub-)component-level results from **existing systems**?
2. Can we increase result **explainability** for (mostly) black-box systems?
3. How may approaches be compared in an in-depth fashion? (**Comparability**)
4. How to properly reproduce existing systems? (**Reproducibility**)

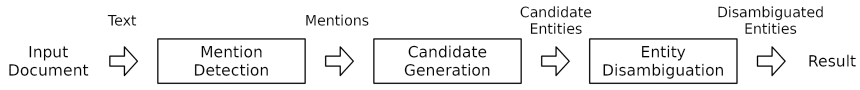
To the best of our knowledge, no execution system attempting to fill the gaps of maximising reusability and comparability, additionally to minimising future development efforts for annotation approaches, exists. As such, we introduce CLiT, our means of simplifying life for and pleasing researchers as well as practitioners in the field of entity linking.

We advance the state-of-the-art by:

1. Introducing novel concepts for EL workflows, including compatibility with existing paradigms;
2. Allowing for nigh-infinite configurability of supported components in complex pipelines;
3. Enabling down-stream processing of annotation results rather than metrics;
4. Improving reusable components from existing systems and ones to come, increasing degree of system support;
5. Providing a knowledge graph agnostic and potentially multi-knowledge graph-supporting annotation service (through *translator*-subcomponents);
6. Defining open exchange protocols based on the Agnos [6] framework, JSON and NIF 2.0 for Mention Detection (MD), Candidate Generation (CG), Entity Disambiguation (ED) as well as *pre- and post-processing subcomponents* acting logically between the aforementioned;
7. Allowing simple introduction of existing systems through RESTful standards.

For further details on CLiT including a demonstration video, we refer interested parties to our Github page (`https://github.com/kmdn/CLiTESWC2021`).

Fig. 1. Classical Pipeline for an EL system. Consisting of mention detection (MD), candidate generation (CG) and entity disambiguation (ED).



2 System Design

Classical Pipeline While EL systems vary in terms of approaches and potential steps within respective pipelines, we identify the most commonly-employed ones as the *classical pipeline*. We use said pipeline as a template for our framework in order to reach compatibility with as many existing systems as possible. In Figure 1, we present our understanding of the functioning of a classical pipeline for a single system.

CLiT Framework Currently, CLiT has integrated six annotators, namely Babelfy [3], DBpediaSpotlight [5], AIDA [4], FOX [9], EntityClassifierEU [2] and OpenTapioca [1] – with more on the way. In order to allow for customised experiences and configurations, in addition to elements pertaining to the *classical pipeline* and entire annotators, we introduce further processing capabilities with the intent of allowing for nigh-infinite combinations of system components (see Fig. 2). We refer to them as *processors* or *subcomponents*, handling post-processing of structures' output from prior tasks, preparing them for being, in turn, potentially further processed by subsequent steps in the chosen workflow. In this paper, we define 4 types of processors: *splitter*s, *combiner*s, *filter*s and *translator*s.

Splitter Allowing for processing of items prior to passing them on to a subsequent step, a splitter is utilised in the case of a *single* stream of data being sent to *multiple* components, potentially warranting specific splitting of data streams (e.g. people-related entities being handled by one system, while another processes movies). This step encompasses both a post-processing step for a prior component, as well as a pre-processing step for a following one. A potential post-processing step may be to filter information from a prior step, such as eliminating superfluous candidate entities or unwanted mentions.

Combiner As a counterpart to a splitter, a *combiner* subcomponent must be utilised in case multiple components were utilised in a prior step and are meant to be consolidated through a variety of possible combination actions (e.g. union and intersection). It combines results from multiple inputs into a single output, passing merged partial results on to a subsequent component.

Filter In order to allow removal of particular sets of items through user-defined rules or dynamic filtering, we introduce a subcomponent capable of processing results on binary classifiers: a *filter*. The truth values evaluated on passed

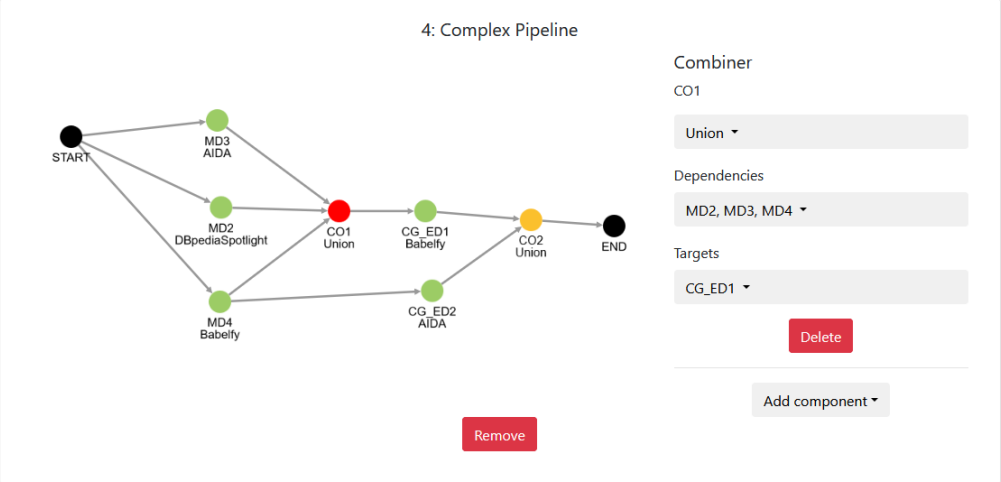


Fig. 2. GUI of CLiT with an example of a *complex* EL pipeline.

partial results define which further outcomes may be detected by a subsequent component or translator.

Translator Enabling seamless use of annotation tools regardless of underlying Knowledge Graph (KG), the translator subcomponent is meant as a processing unit capable of *translating* entities and potentially other features, allowing further inter-system compatibility. It may be employed at any level and succeeded by any (sub-)component due to its ubiquitous characteristics and necessity when working with heterogeneous systems.

3 Conclusion & Future Work

In this paper, we introduced CLiT, a framework for the combination and execution of multiple entity linking approaches, both novel and existing. We show how components classically interact with each other based on a commonly-adopted pipeline and how they may be utilised, as well as extended through our framework. Currently our framework supports six end-to-end entity linking systems for execution in their entirety, as well as in combination with each other. Furthermore, we will introduce semi-automated in-depth analysis features in the future, allowing for collaborative evaluation, yielding a more fine-granular evaluation view on both annotators as well as data sets. Our contributions also increase the ease to train meta learning annotation classifiers with advanced degrees of flexibility and adaptability in relation to textual features.

References

1. Delpeuch, A.: OpenTapioca: Lightweight Entity Linking for Wikidata. In: Proceedings of the 1st Wikidata Workshop co-located with the 19th International Semantic Web Conference. Wikidata'20, vol. 2773. CEUR-WS.org (2020)
2. Dojchinovski, M., Kliegr, T.: Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 654–658. ECML-PKDD'13, Springer (2013)
3. Flati, T., Navigli, R.: Three Birds (in the LLOD Cloud) with One Stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy. In: Proceedings of the Posters and Demos Track of 10th International Conference on Semantic Systems. SEMANTiCS'14, vol. 1224, pp. 10–13. CEUR-WS.org (2014)
4. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust Disambiguation of Named Entities in Text. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 782–792. EMNLP'11, ACL (2011)
5. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings the 7th International Conference on Semantic Systems. pp. 1–8. I-SEMANTICS'11, ACM (2011)
6. Noullet, K.: KG-Agnostic Entity Linking Orchestration. In: Proceedings of the Doctoral Consortium at ISWC 2020 co-located with 19th International Semantic Web Conference. ISWC'20, vol. 2798, pp. 41–48. CEUR-WS.org (2020)
7. Rizzo, G., van Erp, M., Troncy, R.: Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. pp. 4593–4600. LREC'14, European Language Resources Association (ELRA) (2014)
8. Rizzo, G., et al.: NERD: a framework for unifying NERD extraction tools. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics
9. Speck, R., Ngomo, A.N.: Named Entity Recognition using FOX. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track of the 13th International Semantic Web Conference. ISWC'14, vol. 1272, pp. 85–88. CEUR-WS.org (2014)
10. Usbeck, R., Röder, M., Ngomo, A.N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., Wesemann, L.: GERBIL: General Entity Annotator Benchmarking Framework. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1133–1143. WWW'15, ACM (2015)
11. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - A Link Discovery Framework for the Web of Data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web. LDOW'09, vol. 538. CEUR-WS.org (2009)