Contents

List of Figures

List of Tables

Chapter 1

Related Work

Entity linking is a fundamental task in natural language processing that aims to associate references with specific entities from a knowledge base. This can be helpful for facilitating information retrieval, semantic understanding, and knowledge extraction in various domains. In this section, we provide an overview of existing tools for entity linking, with special focus on their application in diverse text domains, the underlying knowledge bases they utilize, and the varied methods employed for entity linking.

Kristianto et al. [6] developed an entity linking method, MEL (Math Entity Linking), to link mathematical expressions from scientific documents to their corresponding Wikipedia articles, especially from the STEM field (Science, Technology, Engineering and Math). The entity linker performs three main tasks, first identifying math expression mentions, then generating candidates for each mention and lastly the disambiguation of the generated candidates. On a dataset of 46 math expressions the approach achieved a precision of 15.22% for highly relevant links and a precision of 36.69% for partially relevant links.

Entity linking is also used in in the field of biomedicine. Chen et al. [2] developed a lightweight neural model for biomedical entity linking. It utilizes an alignment layer with an attention mechanism to capture variations between mentions and entity names, achieving a performance comparable to state-of-the-art BERT-based models. The approach was evaluated on three datasets (ShARe/CLEF, NCBI, ADR), achieving a recall of 97.79%, 94.27% and 96.66% respectively. The accuracies of the base model on all three datasets were 90.10% (ShARe/CLEF), 89.07% (NCBI) and 92.63% (ADR).

Liu et al. [7] investigated the process of entity linking in tweets, addressing the challenges posed by limited information and diverse references to entities. They introduced a collective inference approach, integrating three similarity measures to enhance contextual understanding and address irregular mentions. The approach is evaluated on publicly available datasets, showing that it outperforms established baselines and achieved a precision 0.752%, a recall of 0.675% and a F1-score: 0.711%.

In the field of computer science Yasukawa [12] presented an entity linking technique, designed for computer science education. Its goal is to connect keywords found in university syllabi to corresponding entities in Wikipedia. It has been extensively evaluated and found to be highly effective due to its tailored approach to computer science. The technique improves text comprehension for humans and computers by using curriculum standards and syllabi.

AIDA [13] is an online tool, which can be used for entity detection and linking in text and tables to specific entities in a knowledge base. The system utilizes collective disambiguation by taking prominence of entities, the similarity of context, and coherence among candidates into account. By using knowledge from YAGO2 [5] and employing a graph-based approach, it effectively addresses the challenges of named-entity disambiguation.

TagME [4] is a system for annotating short or poorly composed text fragments with relevant hyperlinks to Wikipedia. This system stands out from other existing systems due to the ability to annotate brief and inadequately constructed texts, including snippets from search-engine results or tweets, making it valuable for tasks relying on structured knowledge from Wikipedia, such as text analysis, clustering and information retrieval. The approach was evaluated on three datasets, one was the IITB dataset, the other two were derived from Wikipedia (Wiki-Disamb30 and Wiki-Annot30). For disambiguation the approach achieved a precision of 91.5%, a recall of 90.9% and F-Measure of 91.2%. The performance of annotators of Wiki-Annot30 was evaluated with annotation and topics metric. For annotation metrics the precision, recall and F-measure were 76.27%, 76.08% and 76.17% repectively and for topics metrics they were 78.41%, 77.48% and 77.94%.

DBpedia Spotlight [8] is a system that connects text documents to Linked Open Data, automatically annotating text with DBpedia URIs. Users can customize annotations based on the DBpedia ontology and quality measures. As an open-source tool, DBpedia Spotlight facilitates interlinking web documents with DBpedia, offering a comprehensive solution identifying and disambiguation references to DBpedia resources. The systems disambiguation was evaluated with five different disambiguation approaches, achieving the best accuracy of 80.52% with a mixed approach with their default sense baseline and TF*ICF scores approaches. The best configuration of DBpedia Spotlight achieved a F1-score of 56% for the annotation evaluation.

OpenTapioca [3] is a specialized Named Entity Linking system designed for Wikidata.

It serves to highlight the unique characteristics of Wikidata in comparison to other knowledge bases like DBpedia and YAGO, with particular emphasis on its edible and multilingual nature. The system sets its focus on the structured data that is collaboratively generated within Wikidata, allowing real-time synchronization and immediate benefits from user edits. OpenTapioca's components encompass local compatibility, semantic similarity, and mapping coherence to classify entities in a context.

Another entity linking tool is Falcon 2.0 [9] that focuses on joint entity and relation linking over Wikidata, with the primary objective to link natural language text to appropriate matches in a Knowledge Graph (KG). The tool uses a background KG created from Wikidata and is based on a rule-based approach that relies on English morphology principles. Falcon 2.0's recognition and linking modules include POS tagging, tokenization, and N-gram tiling for extraction, and candidate list generation, matching, and ranking for linking. The entity and relation linking of the tool was evaluated on three datasets (SimpleQuestion, LC-QuAD 2.0, WEBQSP-WD). On the LC-Quad and SimpleQuestion datasets, Falcon 2.0 was compared to OpenTapioca, outperforming it in precision, recall and F1-score. The tool was also compared to the first Falcon approach on the WEBQSP dataset, achieving better results for precision (0.8% against 0.47%), recall (0.84% against 0.45%) and F-score (0.82% against 0.46%). Relation linking was evaluated on LC-QuAD 2.0 and SimpleQuestion resulting in precision of 0.44%, recall 0f 0.37% and F-score of 0.4% for the first dataset and 0.35% for precision, 0.44% for recall and 0.39% for F-score for the the second dataset.

Wikifier [1] is a tool for annotating documents with relevant concepts sourced from Wikipedia, using a pagerank-based approach. It breaks down the procedure in three steps, recognizing phrases that pertain to Wikipedia concepts, disambiguating these phrases, and ascertaining relevant concepts for the entire document. The method constructs a mention-concepts graph, utilizes pagerank scores to globally disambiguate, and filters out mentions that are open to multiple interactions. The approach depends on heuristics to enhance accuracy by addressing mentions that are prone to ambiguity, and incorporates various extensions for disambiguation. Wikifier was evaluated on a dateset of 1393 news articles and compared to other systems, like AIDA, Waikato, Babelfy, Illinois and DbPedia Spotlight. It achieved a F1-score 0.593, performing slightly worse than AIDA(0.723) but better than the other systems.

FOX [10], a Federated knOwledge eXtraction framework, highlighting its role in Named Entity recognition on unstructured web data. Its primary function is to disambiguate and link named entities to DBpedia, providing accurate and linked entities in various RDF serialization formats. The workflow involves preprocessing input data, recognizing Named entities, linking them to resources with AGDISTIS [11], and converting results to RDF format. The framework is designed to accept various input types, cleans HTML tags, and detect sentences and tokens. Additionally, it employs AGDISTIS for entity disambiguation and linking against DBpedia, supporting multiple serialization formats for the output. The approach was evaluated on five different datasets (News, News* (subset of News dataset), Web, Reuters, All), one time token-based and one time entity-based and is compared to Stanford, Illinois, OpenNLP and Balie. FOX outperforms all other tools, achieving the highest F1-score of 95.23% on the News* dataset (token-based).

Title	*dataset*	*method*	*metric*	*scores*
Entity Linking for Mathematical Expressions in Scientific Documents		mathematical expression extraction, Disambiguation, candidate generation	Precision	0.3669
A Lightweight Neural Model for Biomedical Entity Linking	TAC 2017 ADR	recurrent neural network, bidirectional LSTM, Attention Mechanisms	Recall	0.9779
Entity linking for tweets		collective inference method, mention-mention similarity, mention-entry similarity, entity-entity similarity	Precision	0.752
Entity Linking among Categorized Knowledge Resources for Computer Science Curricula		feature word selection, feature document selection	F-score	0.715
AIDA		mutual information, Named entity disambiguations, graph based approach		

TagMe	Wiki-Disamb30	anchor disambiguation, anchor parsing, anchor pruning	F-score	0.912
DBpedia Spotlight		Aho–Corasick algorithm, Hidden Markov Model, vector space model, Customizable annotation system	F-score	0.56
Falcon 2.0	Webqsp-wd	tokenization, n-gram splitting, Rule-Based Approach, part-of-speech tagging, joint entity and relation linking	F-score	0.82
OpenTapioca		Semantic similarity, linear support classifier, graph based approach, binary classification		
Wikifier		page-rank based wikification, mention-concept graph, global disambiguation	F-score	0.593
FOX	News*	rdf serialization, ensemble learning	F-score	0.9523

Bibliography

[1] Janez Brank, Gregor Leban, and Marko Grobelnik. “Annotating documents with relevant wikipedia concepts”. In: *Proceedings of SiKDD* 472 (2017).

[2] Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. “A lightweight neural model for biomedical entity linking”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 14. 2021, pp. 12657–12665.

[3] Antonin Delpeuch. “Opentapioca: Lightweight entity linking for wikidata”. In: *arXiv preprint arXiv:1904.09131* (2019).

[4] Paolo Ferragina and Ugo Scaiella. “Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 1625–1628.

[5] Johannes Hoffart et al. “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”. In: *Artificial intelligence* 194 (2013), pp. 28–61.

[6] Giovanni Yoko Kristianto, Goran Topić, and Akiko Aizawa. “Entity linking for mathematical expressions in scientific documents”. In: *Digital Libraries: Knowledge, Information, and Data in an Open Access Society: 18th International Conference on Asia-Pacific Digital Libraries, ICADL 2016, Tsukuba, Japan, December 7–9, 2016, Proceedings 18*. Springer. 2016, pp. 144–149.

[7] Xiaohua Liu et al. “Entity linking for tweets”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 1304–1311.

[8] Pablo N Mendes et al. “DBpedia spotlight: shedding light on the web of documents”. In: *Proceedings of the 7th international conference on semantic systems*. 2011, pp. 1–8.

[9] Ahmad Sakor et al. “Falcon 2.0: An entity and relation linking tool over wikidata”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 3141–3148.

[10] René Speck and Axel-Cyrille Ngonga Ngomo. “Named Entity Recognition using FOX.” In: *ISWC (Posters & Demos)*. Citeseer. 2014, pp. 85–88.

[11] Ricardo Usbeck et al. “AGDISTIS–agnostic disambiguation of named entities using linked open data”. In: *ECAI 2014*. IOS Press, 2014, pp. 1113–1114.

[12] Michiko Yasukawa. “Entity Linking among Categorized Knowledge Resources for Computer Science Curricula”. In: *IIAI Letters on Institutional Research* 3 (2023).

[13] Mohamed Amir Yosef et al. “Aida: An online tool for accurate disambiguation of named entities in text and tables”. In: *Proceedings of the VLDB Endowment* 4.12 (2011), pp. 1450–1453.