WikiGUM: Exhaustive Entity Linking for Wikification in 12 Genres

Jessica Lin
Department of Linguistics
Georgetown University
yl1290@georgetown.edu

Amir Zeldes
Department of Linguistics
Georgetown University
amir.zeldes@georgetown.edu

Abstract

Previous work on Entity Linking has focused on resources targeting non-nested proper named entity mentions, often in data from Wikipedia, i.e. Wikification. In this paper, we present and evaluate WikiGUM, a fully wikified dataset, covering all mentions of named entities, including their non-named and pronominal mentions, as well as mentions nested within other mentions. The dataset covers a broad range of 12 written and spoken genres, most of which have not been included in Entity Linking efforts to date, leading to poor performance by a pretrained SOTA system in our evaluation. The availability of a variety of other annotations for the same data also enables further research on entities in context.

1 Introduction

Entity linking (EL) involves identifying entities within a text and subsequently linking their mentions to a knowledge base or table of authorities. The former step is often referred to as Named Entity Recognition (NER) and the latter may also be referred to as entity disambiguation. In this study, we will focus on the latter task by following the popular approach of mapping named entities to Wikipedia entities (Milne and Witten, 2008; Shnayderman et al., 2019), i.e. Wikification.

Wikification is the task of adding links to Wikipedia pages to mentions of named entities in a written or spoken text. This task supports Natural Language Understanding in downstream tasks such as question answering, summarization, and relation extraction. However the scope and structure of EL depends heavily on datasets which are either automatically derived from hyperlinked text and thus suffer some limitations, or are created via human annotation, a time-consuming and expensive task. Despite numerous existing EL datasets (Cucerzan, 2007; Ji et al., 2015; Kulkarni et al., 2009; Milne

and Witten, 2008; Ratinov et al., 2011), few have attempted to capture *nested* entity structure, as in Figure 1, which never occurs in hyperlinks, which cannot be nested.
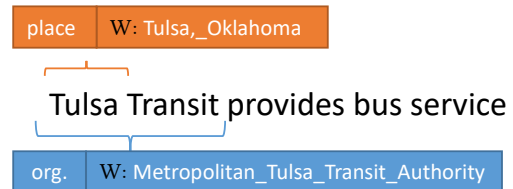

Figure 1: Nested entity linking.

Instead, annotations have focused on flat mention structure from popular online sources, leaving out important information in nested entities that can be useful for downstream tasks. Closest to the resource presented here is the Nested Named Entities (NNE) dataset (Ringland et al., 2019), which is a large, manually-annotated, nested named entity dataset over English newswire, however, it does not include entity linking. Although NNE includes fine-grained semantic information in nested entity types, it is not linked to any identifiers (e.g. a Wikipedia page). Furthermore, even if used for mention recognition, the data is not ideal for testing on diverse genres, as NNE solely covers news text. We also note other datasets capturing some nested entity structure, such as the Abstract Meaning Representation (AMR) corpus (Banarescu et al., 2013), which includes compositional nesting e.g. in possessives such as *Toronto's international airport*, composed of a city and an airport. However, since AMR is not word-aligned to text, even those nested entities that are covered are not aligned to their textual position.

In this paper, we present and evaluate a gold standard wikified dataset, called WikiGUM, in which named and non-named entities have been annotated manually. WikiGUM is based on the existing GUM dataset (Georgetown University Multilayer corpus,

Zeldes 2017), and goes beyond other EL corpora, in covering *all* mentions of named entities (NEs), including non-named and pronominal mentions, as well as nested mentions, for 12 genres of English text. WikiGUM also enables assessment of EL annotations by highlighting challenges that are common in our dataset, and reveals the relatively poor coverage of state-of-the-art NLP systems for EL in diverse genres (Section 4). Taken together, we aim to facilitate new research on nested NER and EL, to promote recognition of all NE mentions and a deeper understanding of the hierarchical structure of entities in text.

2 WikiGUM

The underying GUM (Zeldes, 2017) corpus is a manually annotated dataset with multiple layers, including POS tagging (Penn tags, CLAWS5, Universal POS), sentence types (e.g. declarative, imperative, yes/no question), UD dependency trees (Nivre et al., 2016), coreference resolution (including bridging anaphora and split antecedents), and RST discourse parses (Mann and Thompson, 1988). Data covers 12 genres: academic, biographies, conversation, fiction, forums, how-to, interviews, news, speeches, textbooks, travel and vlogs.

WikiGUM adds a layer of Wikipedia identifiers to all NEs in GUM, which are identified automatically by having the gold PTB POS tag NNP(S) for their syntactic head, based on gold syntax trees (for some resulting issues, see below), as well as non-named mentions coreferring to them based on coreference annotations. Since GUM is expanded by students in classroom annotation every year, and we plan to continue adding Wikification in the future, no closed or pre-prepared ontology is applied to the Wiki identifiers, making the task simpler for student annotators who only need to find a corresponding Wikipedia article.

That said, the existing 10 entity types in GUM[1] mean that our EL benefits from the same categorization scheme as a rough ontology, and the availability of semantic information from WikiData means that many relationships between entities can be explored. All referential NPs, including pronouns and even clauses (if they co-refer with a named entity based on GUM's coreference annotations, for example movie titles), were selected as markables for annotation. Note that nested markables are always

___

[1]Types: PERSON, PLACE, ORG, ANIMAL, PLANT, EVENT, TIME, SUBSTANCE, ABSTRACT and inanimate OBJECT.

included, for example:

(1)   [*the airport in* [*Cuba*]place]place

Our general guideline for entity linking is that NEs, including pronominal and non-named mentions, were manually linked to the corresponding Wikipedia article whenever one exists, using the version controlled online editor GitDox (Zhang and Zeldes, 2017). For example:

(2)   *Kim likes* [*The Terminator*]abstract. [*This movie*]abstract *is her favorite.*

In this example, the span *This movie* should also be linked to the Wikipedia page that refers to *The Terminator (the movie)*. Statistics on WikiGUM, which is freely available under the same Creative Commons license as GUM, are shown in Table 1.

Although the basic Wikification task is fairly straightforward, some ambiguous/tricky cases during annotation included:

- **Generic terms**: some capitalized common nouns that have Wikipedia links appear within NEs, and are tagged NNP(S), but do not correspond to named entities. For example, *Oil* is incorrectly proposed as a NE due to capitalization within the NE *the Oil Capital of the World* (referring to Tulsa, OK) and due to the POS tag NNP. It can be tempting to link 'oil' as a NE candidate to the Wikipedia article 'Petroleum'. However in context, 'oil' is a generic, non-named modifier to 'Capital', and should not be linked as a NE. Annotators should be mindful of context of terms tagged NNP(S) within NEs, rather than linking any NNP span.

- **Subset of entity with the same type**: a common type of ambiguity for PLACE entities arises when names are reused in different countries, regions, cities or villages. For example, terms like 'North', 'South', 'East', and 'West' as a subset of a region are hard to disambiguate, and they are common in street names in North America. In this case, annotators must look at the broader context and carefully check whether the entity refers to a subset or not, for example cities and their metropolitan areas, streets with and without cardinal directions, or other parts of cities which sometimes have separate Wikipedia entries.

Text Types	Source	Documents	# of NE Mentions	# of Nested Wikified Mentions	Total Mentions	Tokens
Interviews	Wikinews	19	1,146	107	5,204	18,037
News stories	Wikinews	21	1,221	217	4,130	14,094
Travel guides	Wikivoyage	17	1,327	174	4,087	14,955
How-to guides	WikiHow	19	94	8	4,469	16,920
Academic writing	various	16	329	48	4,486	15,110
Biographies	Wikipedia	20	2,450	413	5,763	17,951
Fiction	various	18	195	9	4,737	16,307
Forum discussions	Reddit	18	196	2	4,530	16,286
Conversations	UCSB corpus	5	31	0	1,477	5,698
Political speeches	various	5	316	32	1,423	4,831
CC Vlogs	YouTube	5	39	1	1,355	5,180
Textbooks	OpenStax	5	188	21	1,507	5,376
Total		168	7,352	1,032	43,168	150,745

Table 1: Statistics on WikiGUM

- **Distinct links for identical mentions**: It is sometimes hard for annotators to realize that an entity string has several EL variants. This happens often in abbreviations, which may be labelled with the wrong entity type. For example, 'JFK' can be a person's name (the 35th US President) or a place name (JFK Airport in New York), depending on context. We instructed annotators to prioritize the existing entity type annotation: if JFK is tagged as a place, it is linked to the article about the airport. Another common issue affects ancient place names which do not exist nowadays, resulting in difficulty for EL. For instance, *Jorvik* is the viking name of York, and was therefore linked to the closest equivalent article, 'Scandinavian York' rather than 'York'. In other cases, we relied on the coreference annotations to establish equivalence: for example *England's City of Festivals* was labeled as coreferring with *York*, and was therefore considered equivalent to York for EL purposes.

- **Lack of background knowledge**: In some cases context alone cannot help annotators decide on the right sense of an entity, especially in academic texts, but also in discussion forums. Academic articles often assume readers have detailed knowledge of the topic and thus provide little context for the target entity. For example, 'Su' in 'Su et al. 2016' is a named entity, but it may be difficult to know whether there is a corresponding Wikipedia article based solely on the author's name.

3 Related work

Table 2 compares WikiGUM to other EL corpora. Most current EL datasets are based on newswire text, overlooking the impact of genre on EL – for example, Dai (2018) notes that the biomedical domain involves complex and unique entity mentions. As EL datasets are developed for evaluation of EL systems, out-of-domain data could create challenges for conventional tools. Furthermore, most previous work (Cucerzan, 2007; Ji et al., 2015; Kulkarni et al., 2009; Milne and Witten, 2008; Ratinov et al., 2011) has focused on identifying and classifying atomic, flat mention structures, leaving out the semantic information available in nested mentions.

As shown in Table 2, most datasets do not contain nested entity linking annotations, with ACE2004 being the exception (Ratinov et al., 2011). Unfortunately, no dataset covers all mentions of named entities, including their non-named mentions ('the same airport', or 'it'). As mentioned above, we do see some research on nested entity structure (Glavaš and Šnajder, 2014; Hong et al., 2016), for example the NNE corpus (Ringland et al., 2019) contains fine-grained semantic information including e.g. the category CITY nesting a STATE, which could easily be used for EL. However they are not disambiguated or linked to a table of authorities, in addition to excluding non-named mentions of the same entities. WikiGUM thus differs from previous EL datasets and is rich in terms of both genre and entity structure, as well as being among the larger available datasets as shown in the Table.

4 Evaluation

In this section we evaluate inter-annotator agreement, as well as the extent to which existing Wikification technology already captures the information in WikiGUM.

Dataset	Paper	# of documents	# of NEs	# of genres	NE/N	Pronouns	Nested Entities
WikiGUM	–	168	7,352	12	NE&N	✓	✓
ACE2004	(Ratinov et al., 2011)	36	256	1	NE		✓
AIDA-A	(Hoffart et al., 2011)	216	5,917	1	NE		
AIDA-B	(Hoffart et al., 2011)	231	5,616	1	NE		
AQUAINT	(Milne and Witten, 2008)	50	727	1	NE&N		
Derczynski	(Derczynski et al., 2015)	182	210	1	NE		
IITB	(Kulkarni et al., 2009)	107	17,200	1	NE		
KORE50	(Hoffart et al., 2012)	50	148	1	NE		
MSNBC	(Cucerzan, 2007)	20	656	1	NE		
n3-RSS-500	(Röder et al., 2014)	500	1,000	1	NE		
n3-Reuters-128	(Röder et al., 2014)	128	880	1	NE		
OKE2015	(Nuzzolese et al., 2015)	not specified	718	1	NE&Roles		
OKE2016	(Nuzzolese et al., 2016)	not specified	940	1	NE&Roles		

Table 2: English EL datasets. NE/N indicates whether only named entities (NE) or also common nouns (N) are included. Note that ACE 2004 is a subset of the documents used in the ACE 2004 coreference documents.

Inter-annotator agreement Measuring agreement for Wikification involves two main complementary aspects: span detection and Wikification (including the decision whether to link an entity and to what). Since GUM already contains mention boundaries and named/non-named status, we focus on the latter task, measuring linking agreement. To calculate agreement, we carried out an inter-annotator agreement experiment by double annotating 3,103 tokens of corpus data containing 237 entities after adjudication, about 3% of the data. We compute both Cohen's Kappa and simple percent agreement (percentage of exact match), shown in Table 3. Note that computing Cohen's Kappa here is somewhat artificial, as in the real world there is an (almost) unlimited space of possible Wikipedia identifiers. For simplicity, we define the space of possible links as the union of any values annotators used in this subset, meaning that any link chosen at any point by any annotator is considered a possible value for the annotation, and any disagreement is penalized by the metric.[2]

Metric	Score
Agreement	0.8903
Cohen's κ	0.8782

Table 3: Results of the inter-annotator agreement experiment

Results in Table 3 show that agreement is far beyond chance, with Kappa=.87 and simple agreement of .89. While this indicates very good agreement, raters did disagree on ambiguous cases, which is worth discussing. A major source of disagreement involves linking the same entity string to distinct but related identifiers, i.e. the **name variants** issue highlighted in Section 2. This is often due to lack of context information: for example, in the sentence "*CC makes things more complex*", the mention *CC* could be linked to *Creative Commons license* (public copyright license) or *Creative Commons* (organization that produced the Creative Commons license). In this case, broader context and reasoning are required to make consistent decisions. Another example is the sentence "*According to the Arts and Humanities Citation Index Professor Chomsky is the eighth most cited scholar of all time.*", the mention *Arts* is not a NE and should not be linked, but was linked by one annotator to "*The arts*", which has a linkable article and can easily be confused with a named concept of sorts.

NLP coverage To evaluate the usefulness of WikiGUM beyond existing resources, we test a recent SOTA pretrained end-to-end neural system (*e2e*, Kolitsas et al. 2018) on the test set and compare it to a baseline strategy. Our baseline system uses a neural constituent parser (Mrini et al., 2020) to identify predicted noun phrases and simply checks the exact string of every phrase headed by a proper noun to see if it has a Wikipedia article (using the Python library `wikipedia`). Since the SOTA system cannot identify nested mentions, but we do not know which one of two nested mentions it might identify (the bigger or smaller one), we

---

[2]An anonymous reviewer has asked whether this means that search ambiguity and an overwhelming number of options impacted our process: this is certainly true, and somewhat inevitable given that annotators were unrestricted in the identifiers they could choose from Wikipedia. However the high level of absolute agreement suggests that in practice annotators were surprisingly internally consistent.

evaluate in multiple scenarios: counting all mentions, only unnested mentions, and, since the system cannot identify pronouns, with and without them.

	data	P	R	F1	links
e2e	all	0.398	0.192	0.259	827
	-pron	0.441	0.259	**0.327**	677
	-nest	0.363	0.203	0.260	713
	-both	0.363	0.253	0.298	573
baseline	all	0.480	0.182	0.264	827
	-pron	0.480	0.223	**0.304**	677
	-nest	0.363	0.203	0.260	713
	-both	0.391	0.214	0.277	573

Table 4: Baseline and e2e results on WikiGUM test set.

The results in 4 show that *e2e*, even when trained on the largest available Wikification dataset (AIDA, ≈1,000 documents, 18K links) does not generalize well to the domains found in our corpus, barely outperforming a naive lookup baseline. Comparing the scenarios, we see that best performance for both systems is achieved when removing pronouns, which is unsurprising since neither strategy can be expected to link them. However removing nested mentions does not result in higher scores: this is because some common targets, such as places, are often nested in larger names (organizations, office-holders), and removing them disrupts score gains from their correct identification. Nevertheless, the most lenient possible evaluations are in the low 30s, as opposed to a score of 82.6 on AIDA (Kolitsas et al., 2018, 524). This suggests that, unsurprisingly, the corpus covers a range of entities and contexts that are under- or unrepresented in previous benchmarks.

5 Conclusion

This paper presented WikiGUM, the first exhaustive Wikification dataset for named entity linking, including nested and pronominal mentions in 12 genres of English text. Our evaluation suggests a high level of agreement, as well as coverage for a significant amount of entities not retrieved by a SOTA neural linking system. We hope that this dataset will enable further research on entity linking and increase coverage for all types of linkable named entities across a broad spectrum of genres, both spoken and written.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 708–716.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Goran Glavaš and Jan Šnajder. 2014. Constructing coherent event hierarchies from news stories. In *Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*, pages 34–38.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 545–554.

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 the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 1–6.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *TAC*.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

David Milne and Ian H Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 2015. Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*, pages 3–15, Cham. Springer International Publishing.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Robert Meusel, and Heiko Paulheim. 2016. The second open knowledge extraction challenge. In *Semantic Web Evaluation Challenge*, pages 3–16. Springer.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1375–1384.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. *arXiv preprint arXiv:1906.01359*.

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3-A collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *LREC*, pages 3529–3533.

Ilya Shnayderman, Liat Ein-Dor, Yosi Mass, Alon Halfon, Benjamin Sznajder, Artem Spector, Yoav Katz, Dafna Sheinwald, Ranit Aharonov, and Noam Slonim. 2019. Fast end-to-end wikification. *arXiv preprint arXiv:1908.06785*.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Shuo Zhang and Amir Zeldes. 2017. GitDOX: A linked version controlled online XML editor for manuscript transcription. In *FLAIRS Conference*, pages 619–623.