Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing

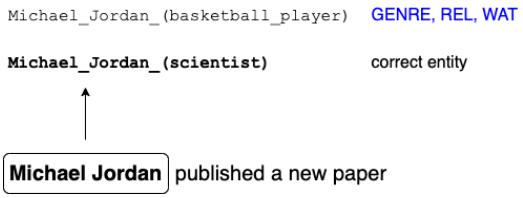
Vera Provatorova, Samarth Bhargav, Svitlana Vakulenko, Evangelos Kanoulas
University of Amsterdam, Amsterdam, The Netherlands
{v.provatorova, s.bhargav, s.vakulenko, e.kanoulas}@uva.nl

Abstract

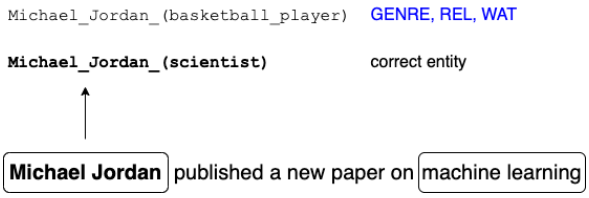
Entity disambiguation (ED) is the last step of entity linking (EL), when candidate entities are reranked according to the context they appear in. All datasets for training and evaluating models for EL consist of convenience samples, such as news articles and tweets, that propagate the prior probability bias of the entity distribution towards more frequently occurring entities. It was previously shown that performance of EL systems on such datasets is overestimated, since it is possible to obtain higher accuracy scores by merely learning the prior. To provide a more adequate evaluation benchmark, we introduce the ShadowLink dataset, which includes 16K short text snippets annotated with entity mentions. We evaluate and report the performance of several popular EL systems on the ShadowLink benchmark. The results show a considerable difference in accuracy between common and uncommon ambiguous entities that require disambiguation, for all of the EL systems under evaluation, demonstrating the effects of prior probability bias and entity overshadowing.

1 Introduction

The task of entity linking (EL) refers to finding named entity mentions in unstructured documents and matching them with the corresponding entries in a structured knowledge graph (Milne and Witten, 2008; Oliveira et al., 2021). This matching is usually done using the surface form of an entity, which is a text label assigned to an entity in the knowledge graph (van Hulst et al., 2020). Some mentions may have several possible matches: for example, "Michael Jordan" may refer either to a well-known scientist or the basketball player, since they share the same surface form. Such mentions are ambiguous and require an additional step of entity disambiguation (ED), which is conditioned on the context in which the mentions appear in the text, to be linked correctly. Following van Erp and Groth (2020) we refer to a set of entities that share the same surface form as an *entity space*.



(a) "Michael Jordan" (scientist) is overshadowed by "Michael Jordan" (basketball player).



(b) Even with more relevant context, overshadowing persists.

Figure 1: An example of entity overshadowing. The correct entity is ranked lower by the EL systems (indicated in blue) than the more common one.

To decide which of the possible matches is the correct one, an ED algorithm typically relies on: (1) contextual similarity, which is derived from the document in which the mention appears, indicating the *relatedness* of the candidate entity to the document content, and (2) entity importance, which is the prior probability of encountering the candidate entity irrespective of the document content, indicating its *commonness* (Milne and Witten, 2008; Ferragina and Scaiella, 2012; van Hulst et al., 2020).

The standard datasets currently used for training and evaluating ED models, such as AIDA-CoNLL (Hoffart et al., 2011) and WikiDisamb30 (Ferragina and Scaiella, 2012), are collected by randomly sampling from common data sources, such as news articles and tweets. Therefore, they are expected to mirror the probability distribution

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10501–10510
November 7–11, 2021.

with which the entities occur, thereby favouring more frequent entities (head entities) (Ilievski et al., 2018). From these considerations, we conjecture that the performance of existing EL algorithms on the ED task is overestimated. We set out to explore this effect in more detail by introducing a new dataset for ED evaluation, in which the entity distribution differs from the one typically used for training ED algorithms.

We perform a systematic study focusing on a particular phenomenon we refer to as *entity overshadowing*. Specifically, we define an entity e_1 as overshadowing an entity e_2 if two conditions are met: (1) e_1 and e_2 belong to the same entity space S, i.e., share the same surface form and, therefore, can be confused with each other outside of the local context; (2) e_1 is more common than e_2 in some corresponding background corpus (e.g. the Web), i.e., it has a higher prior probability $P(e_1) > P(e_2)$.

For example, e_1 = "Michael Jordan" (basketball player) overshadows e_2 = "Michael Jordan" (scientist) because $P(e_1) > P(e_2)$ in a typical dataset sampled from the Web. We use an unambiguous text sample that contains this mention to evaluate three popular state-of-the-art EL systems, GENRE (De Cao et al., 2020), REL (van Hulst et al., 2020), and WAT (Piccinno and Ferragina, 2014), and empirically verify that the overshadowing effect that we hypothesized, indeed, takes place (see Fig. 1a). Even when more information is added to the local context, including the directly related entities that were correctly recognised by the system ("machine learning"), the ED components still fail to recognise the overshadowed entity (see Fig. 1b).

The concept of overshadowed entities introduced in this paper is related to long-tail entities (Ilievski et al., 2018). However, these two concepts are distinct: a long-tail entity may be unambiguous and therefore not overshadowed, while an overshadowed entity may still be too popular to be considered a long-tail one.

To systematically evaluate the phenomenon of entity overshadowing that we have identified, we introduce a new dataset, called ShadowLink. ShadowLink contains groups of entities that belong to the same entity space. Following van Erp and Groth (2020), we use Wikipedia disambiguation pages to collect entity spaces. Disambiguation pages group entities that often share the same surface form and may be confused with each other. We then follow the links in the Wikipedia disambiguation pages to the individual (entity) Wikipedia pages to extract text snippets in which each of the ambiguous entities occur.

Note that we do not extract the text from these Wikipedia pages directly, since pre-trained language models such as BERT (typically used in state-of-the-art ED systems) also use Wikipedia as a training corpus, and can learn certain biases as well. Instead, we parse external web pages that are often linked at the end of a Wikipedia page as references. This data collection approach helps us to minimise the possible overlap between the test and training corpus.

Thereby, every entity in ShadowLink is annotated with a link to at least one web page in which the entity is mentioned. We then proceed to extract all text snippets in which the corresponding entity mention appears on the page. An extracted text snippet typically consists of the sentence in which the mention occurs.

Next, we use ShadowLink to answer the following research questions:

RQ1: How well can existing ED systems recognise overshadowed entities?

RQ2: How does performance on overshadowed entities compare to long-tail entities?

RQ3: Are ED predictions biased and how can we measure this bias?

Our contribution is twofold: (1) a new dataset for evaluating entity disambiguation performance of EL systems specifically focused on overshadowed entities, and (2) an evaluation of current state-of-the-art algorithms on this dataset, which empirically demonstrates that we correctly identified the type of samples that remain challenging and provide an important direction for future work.

2 The ShadowLink Dataset

This section describes the ShadowLink dataset: its construction process, structure, and statistics.

2.1 Dataset construction

The process of dataset construction consists of 3 steps: (1) collecting entities, (2) retrieving context examples for each entity, and (3) filtering the data based on the validity requirements detailed below.

Collecting entities. Similar to van Erp and Groth (2020), we use Wikipedia disambiguation pages to represent entity spaces. We retrieve a set

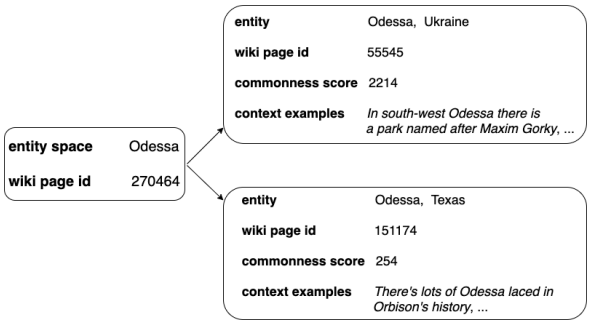


Figure 2: Structure of the ShadowLink dataset

of all Wikipedia disambiguation pages and filter it on the following criteria:

(1) For each disambiguation page (DP), we only include candidate entity pages with names containing the title of the DP as a substring. This step is required to exclude synonyms and redirects.

(2) If at least two candidate pages for the same DP match the criterion described above, then the DP and all its matching candidates are included as a new entity space.

During the first stage of the data collection, 170K out of 316.5K Wikipedia disambiguation pages matched the filtering criteria described above.

Filtering pages by year. To make sure that all pre-trained EL systems we evaluate in our experiments can potentially recognise all of the entities in the dataset, we also exclude pages that are more recent than the Wikipedia dumps used by these systems during training. The oldest dump used by a system in our experiments was the 2016 Wikipedia dump over which TagMe was trained, i.e we excluded all the pages that were created after 2016.

Collecting context examples. To retrieve context examples for each entity, we follow the external links extracted from the references section of the corresponding Wikipedia page and parse them to extract the text snippets which contain the entity mention. Then, every target entity mention is replaced with its corresponding entity space name, yielding an ambiguous entity mention. For example, if we have entities "John Smith" and "Paul Smith" that both belong to the entity space "Smith", then the mentions of both names will be replaced with "Smith". Looking for an entity name and replacing it with the corresponding entity space name (instead of looking for the entity space name in the first place) allowed us to make sure that the text snippets refer to the correct entity. Using this method, however, significantly reduced the number of retrieved snippets, as many of the entity mentions in natural texts do not include the full titles of the entities.

To extract the text snippets, we used a simple greedy algorithm that starts with the mention boundaries and tries to include more text, expanding the boundaries to the left and to the right, until it either covers one sentence on each side, or reaches the end (or beginning) of the document text. Our decision was to use relatively short spans similar to other popular ED benchmarks: WikiDisamb30 (Ferragina and Scaiella, 2012) and KORE50 (Hoffart et al., 2011). Our manual evaluation confirmed that these spans provide sufficient context for entity disambiguation. We also release the full-text of all web pages as part of our dataset, making the context of different lengths available for future experiments.

Commonness score. We estimate the commonness (popularity) of an entity as the number of links pointing to the entity page from other Wikipedia pages, that is, the in-degree of the entity page in the web graph of Wikipedia hyperlinks. Intuitively, this is proportional to the probability of encountering this entity when sampling a page at random. To obtain this metric for all the entities in the dataset, we use the Backlinks MediaWiki API[1].

Quality assurance. We conduct manual evaluation to assess the quality of the dataset and provide the upper bound performance for the ED task. The details of the setup and the results are discussed in Section 3.

2.2 Dataset structure and statistics

The ShadowLink dataset consists of 4 subsets: *Top*, *Shadow*, *Neutral* and *Tail*. The Top, Shadow and Neutral subsets are linked to each other through the shared entity spaces. On the other hand, the Tail subset, which contains (typically unambiguous) long-tail entities, is not connected to the other three through the same entity spaces. Nevertheless, it is collected in a similar way as the other three subsets.

Top and Shadow subsets. The structure of the Top and Shadow subsets is shown in Figure 2. Every entity e belongs to an entity space S_m, derived from the Wikipedia disambiguation pages, where m is an ambigous mention that may refer to any of

[1] https://www.mediawiki.org/wiki/API:Backlinks

the entities in S_m. Every S_m contains at least two entities: one e_{top} and one or more e_{shadow} entities. Every entity $e \in S_m$ is annotated with a link to the corresponding Wikipedia page and provided with context examples. A context example is a text snippet extracted from one of the external pages which contains the mention m , with a length of 25 words on average.

Neutral subset. To quantify the strength of the prior of each ED system, we synthetically generate data points for which the context around an entity mention is not useful for disambiguating that mention. To do that we use 7 hand-crafted templates. An example of such a template is the following: "It was the scarcity that fueled our creativity. This reminded me of m today." For each entity space, we generated 7 random contexts.

Tail subset. To evaluate the performance of ED systems on long-tail but typically not overshadowed entities, we collect an additional set of entities by randomly sampling Wikipedia pages that have a low commonness score ($<= 56$ backlinks)[2].

Context examples for these pages were collected in the same manner as described above. The resulting dataset matches the size and structure of other ShadowLink subsets, containing 904 entities.

The sampling process used to collect this subset follows the existing definition of long-tail entities(Ilievski et al., 2018), and is controlled for popularity but not for ambiguity. The Tail subset serves as a control group for the experiments conducted in our study, showing that the concept of entity overshadowing differs from the previously studied long-tail entity phenomena.

ShadowLink statistics. The dataset statistics across all the subsets are summarised in Table 1. Note that the *Top*, *Shadow* and *Neutral* subsets are grouped around the same entity spaces, while the *Tail* subset is constructed by sampling the same number of non-ambiguous entities. Every entity space contains at least 2 entities, with the mean number of entities per space being 2.63, median 2, and maximum 10. Figure 3 shows the distribution of commonness in the three subsets: Top, Shadow and Tail.

For the experiments we used a smaller subset of ShadowLink, with only one randomly selected shadow entity per entity space and one text snippet per entity. Thus, every subset contained 904 entities, with the total size of 9K text snippets. The rest of the data is left out as a training set and can be used in future experiments.

[2]This threshold is equal to the median number of backlinks in the *Shadow* subset.

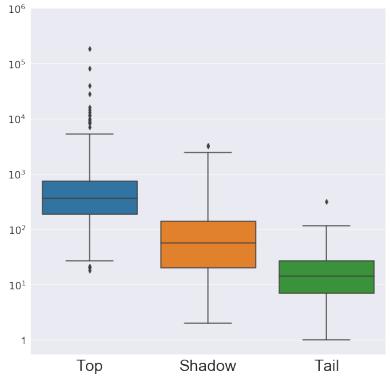


Figure 3: Distribution of the commonness score on the three subsets of ShadowLink.

3 Manual Evaluation

We perform manual evaluation of a random sample from ShadowLink to assess its quality, with the goal of ensuring that the extracted text snippets provide context sufficient for disambiguation. Human performance also sets the skyline for automated approaches on this dataset. In the following subsections, we describe the evaluation setup and the results of the manual evaluation.

3.1 Manual evaluation setup

We conduct a manual evaluation to assess the quality of the dataset and evaluate how well human annotators can disambiguate overshadowed entities. A sample of 91 randomly selected dataset entries was presented to two annotators, who examined the entries independently. For each entry, the annotators were presented with a text snippet containing an ambiguous entity mention m, and two entities, *Top* and *Shadow*, from the same entity space S_m, where one of the two entities was the correct answer. The annotators were instructed to either indicate the correct entity or mark the text snippet as ambiguous, which indicates that the provided context is not sufficient for the disambiguation decision to be made. Note, however, that the commonness scores were not displayed to the annotators.

3.2 Results of the manual evaluation

We used Cohen's kappa coefficient to evaluate the inter-annotator agreement (Bobicev and Sokolova, 2017) on all entries reviewed by the annotators.

Subset	# Entity Spaces	# Entities	# Text Snippets	Avg. # Words	Avg. # Sentences
Top	904	904	2K	29.25	1.11
Shadow	904	1.5K	6K	28.97	1.11
Neutral	904	-	6K	14.83	1.87
Tail	-	904	2K	28.94	1.10

Table 1: Datasets statistics across all the subsets of ShadowLink. The average number of words and sentences were calculated per text snippet extracted from the corresponding web page.

	Shadow	Top
	P = R = F	P = R = F
Annotator 1	0.973	0.973
Annotator 2	0.950	0.919
Average	0.963	0.946

Table 2: Results of the manual annotations.

The value of the coefficient is 0.845, indicative of strong agreement. Next, we discarded the samples labelled as ambiguous by at least one of the annotators. The resulting dataset included 77 entries out of 91, which shows that 85% of the context examples were sufficient for making ED decisions. These unambiguous entries were split into two subsets, resulting in the 37 top-entities and 40 shadow-entities. We then discarded 3 randomly selected shadow-entities to achieve the same size of the two subsets, and used these subsets to evaluate the performance of manual ED for the top- and shadow-entities separately. The averaged F-score of the two annotators is 0.95 on the top-entities and 0.96 on the shadow-entities. The detailed results of the evaluation are shown in Table 2.

The results of manual evaluation show that (1) a majority of samples (85%) in ShadowLink are suitable for ED evaluation, i.e., automatically extracted snippets provide sufficient context for correct disambiguation; (2) human annotators can correctly disambiguate entities regardless of their commonness. Therefore, the performance of an automatic system that only depends on context is only bound by the 15% of the cases for which the context is not helpful. This bound can be further elevated if longer contexts are considered. Experiments on longer contexts are possible using the ShadowLink dataset[3] but we leave it for future work.

In the next section, we report and analyse the results produced by state-of-the-art systems on ShadowLink.

[3]We have crawled the full articles, and will be released as part of the ShadowLink datset.

4 Benchmark Experiments

In this section, we describe the benchmark experiments designed to evaluate the baseline systems' performance on the ShadowLink dataset. For these experiments, we created a subset of the original dataset by sampling only one of the shadow entities at random to make the number of *Top* and *Shadow* equal. Note that in our task setup the model's predictions are not restricted to the top versus shadow entity binary decision. The model can predict any entity from the same or different entity space. We describe the experimental setup in Section 4.1, report the benchmarking results and analyse them in more detail in Section 4.2.

4.1 Evaluation setup

To answer the first two research questions (RQ1 & RQ2), we compare the performance of eight entity linking systems on the ShadowLink dataset. We used the GERBIL framework (Röder et al., 2018) for six of the baselines (AGDISTIS/MAG, AIDA, DBpedia Spotlight, FOX, TagMe 2 and WAT)[4] under the D2KB experimental setup[5]. We also performed an evaluation with the same setup using GENRE and REL, two novel state-of-the-art systems not available in GERBIL. We used micro-averaged precision, recall and F-score as evaluation metrics.

To answer the last research question (RQ3), we

[4]These six systems were the ones available on GERBIL at the time of our experiments.

[5]In the D2KB setup, the systems are provided with correct mention boundaries to evaluate the disambiguation step of entity linking.

want to verify whether the baseline systems utilise context or simply rely on their priors to make the predictions. To this end, we compare the predictions made on the *Top*, *Shadow* and *Neutral* subsets. We used the predictions made on the *Neutral* subset as an indication of priors. That is, for each entity space, we generate context for the *Neutral* subset by using the same 7 random sentences as templates. The context was generated as neutral, i.e., it is not useful for the disambiguation task by design. Therefore, we considered the predictions for such neutral contexts to exhibit the default priors of an EL system for the given entity space. We can then compare these prediction to the predictions on the original examples from the *Top* and *Shadow* subsets. If the entity predicted for non-neutral context differs from the prediction made for the neutral context, we consider that the model updated its default prediction (prior) based on the local context. We performed this type of analysis to examine the predictions of the best-performing systems in our experiments: REL, GENRE, AIDA and WAT.

4.2 Benchmark Results

This section presents the results of our experiments and summarizes the answers to the research questions introduced in Section 1.

RQ1: *How well can existing ED systems recognise overshadowed entities?*

Table 3 shows the evaluation results across the subsets of ShadowLink. All systems achieve the lowest scores on the *Shadow* subset, with the maximum F-score of 0.35 achieved by AIDA. While REL and GENRE ourperform WAT on several existing datasets (van Hulst et al., 2020; De Cao et al., 2020), their results on ShadowLink are considerably lower than the results of WAT. The difference in the results on *Top* and *Shadow* entities indicates that EL predictions are biased towards more common entities.

RQ2: *How does the performance on overshadowed entities compare to long-tail entities?*

All systems show the highest precision on the *Tail* subset, i.e., they achieve much higher performance on the less ambiguous long-tail entities, compared to both top and overshadowed entities in ShadowLink. These results indicate that the main challenge in EL is the combination of ambiguity and uncommonness, while uncommon but non-ambiguous entities are relatively easy to resolve.

These findings are also consistent with Ilievski et al. (2018), who suggest that rare and ambiguous entities constitute the hardest cases for the EL task. In this study, we showed that such overshadowed entities indeed consititute a major challenge for the state-of-the-art systems and that ShadowLink provides a suitable benchmark for their evaluation.

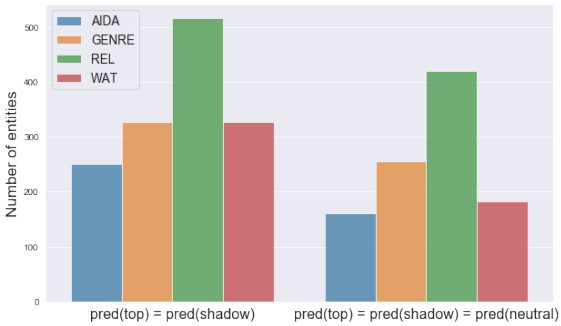


Figure 4: The degree of overshadowing (left) and prior bias (right) for each of the EL systems.

RQ3: *Are ED predictions biased and how can we measure this?*

Our experiments show that all systems under evaluation are often insensitive to the context change, i.e., the systems are actually unable to exploit local context for entity disambiguation but solely rely on their priors learned from the data. The error analysis results presented in Table 4 indicate that the majority of correct answers on the *Top* dataset coincide with the predictions observed on the *Neutral* subset. On the *Shadow* subset, opposite is the case: most of the errors are due to priors, and most of the correct predictions differ from them.

Figure 4 shows the number of cases in which overshadowing occurs for each of the systems, i.e., when the model's prediction remains the same for both *Top* and *Shadow* mentions. We see that this effect correlates with the number of cases in which the prediction of the system does not change regardless of the context, i.e., also for the *Neutral* context the prediction of the system remains the same. This observation confirms our initial hypothesis about the phenomena: the more common entities not only overshadow the less common ones but they are also used as the default predictions made completely independent of the given context, which we call the system priors.

Figure 4 shows that among the four best EL systems, REL is the most prone to overshadowing and

Baseline	Shadow			Top			Tail		
	P	R	F	P	R	F	P	R	F
AGDISTIS/MAG (Usbeck et al., 2014)	0.14	0.14	0.14	0.25	0.25	0.25	0.79	0.79	0.79
AIDA (Yosef et al., 2011)	**0.40**	**0.31**	**0.35**	0.62	0.50	0.56	0.92	0.53	0.67
DBpedia Spotlight (Mendes et al., 2011)	0.16	0.10	0.12	0.41	0.25	0.31	**0.97**	0.11	0.19
FOX (Speck and Ngomo, 2014)	0.17	0.07	0.10	0.29	0.14	0.19	0.82	0.35	0.49
TagMe 2 (Ferragina and Scaiella, 2010)	0.34	0.25	0.29	0.69	0.49	**0.57**	0.95	0.74	0.83
WAT (Piccinno and Ferragina, 2014)	**0.40**	0.19	0.26	**0.72**	0.39	0.51	0.95	0.49	0.65
GENRE (De Cao et al., 2020)	0.26	0.26	0.26	0.42	0.42	0.42	0.93	**0.93**	**0.93**
REL (van Hulst et al., 2020)	0.21	0.21	0.21	0.54	**0.54**	0.54	0.91	0.91	0.91

Table 3: Benchmark evaluation results on the ShadowLink subsets.

	Top					Shadow				
	pred = prior		pred ≠ prior		NIL	pred = prior		pred ≠ prior		NIL
	correct	wrong	correct	wrong		correct	wrong	correct	wrong	
AIDA	15.5	12.9	35.0	28.0	8.6	4.8	20.0	28.2	39.3	7.9
WAT	32.1	11.7	16.4	12.4	27.4	4.2	31.8	20.9	15.6	27.5
GENRE	15.6	32.4	26.7	26.3	0.0	2.2	43.6	23.9	30.3	0.0
REL	32.3	24.8	21.4	21.0	0.6	7.7	50.9	12.9	27.9	0.6

Table 4: Error analysis, which shows the percentage of errors and correct predictions that either coincide (pred=prior) or differ (pred≠prior) from the predictions made for the neutral contexts, which we consider as predictions with the highest prior probability.

prior bias. This also explains its poor performance on the $Shadow$ subset in comparison with the high performance demonstrated on $Tail$. AIDA and WAT appear to be more sensitive to the local context, which allows them to achieve better results on the overshadowed entities in comparison to both GENRE and REL. Moreover, AIDA, which outperforms all other systems on the *Shadow* subset, turns out to be the least affected by the overshadowing phenomena. These results indicate that the main reason behind the poor ED performance on overshadowed entities is due to systems overrelying on the prior bias and failing to incorporate contextual information.

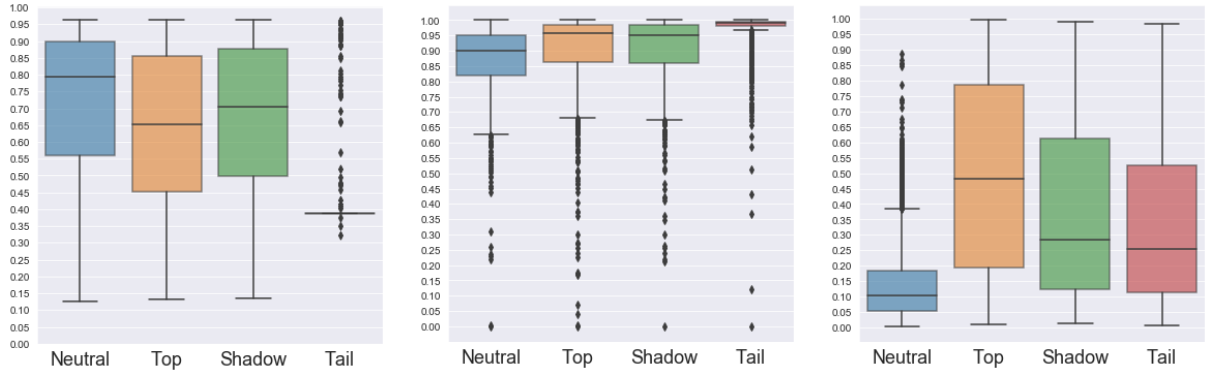
Lastly, we also look at the confidence scores for each of the subsets to check if they can be used as an additional indicator (see Figure 5). Interestingly, the systems have very different distributions of their confidence scores. For example, WAT has lower confidence when given neutral samples, which can be used to detect context ambiguity and filter out such samples. However, this approach can not be used for REL's and GENRE's predictions[6].

[6]GENRE's confidence scores were rescaled before the comparison.

5 Related Work

Datasets for ED evaluation. Evaluation of ED performance was on the research radar for several years, and many benchmark datasets were proposed to date (Hachey et al., 2013; Röder et al., 2018; Ehrmann et al., 2020). Among the most popular ones are AIDA-CONLL (Hoffart et al., 2011), which consists of 1.4K annotated news articles with 27.8 entity mentions; AQUAINT dataset (Milne and Witten, 2008) with 50 news articles and 727 mentions; MSNBC (Cucerzan, 2007) with 20 news articles and 656 mentions. However, the standard benchmarks used for ED evaluation do not reflect the challenges that are often encountered in practice, such as limited context, long-tail, emerging and complex entities (Meng et al., 2021).

Guo and Barbosa (2018) construct two datasets by sampling hard ED examples from Wikipedia and ClueWeb corpora on which a simple baseline using priors does not succeed. Their experiments show that this prior-based baseline achieves a high performance, which also indicates the need for more challenging evaluation datasets. ShadowLink aims to close this gap. In this work, we focus specifically



(a) REL (b) GENRE (c) WAT

Figure 5: Distribution of confidence scores on all subsets of ShadowLink.

Aspect	ShadowLink	WikiDisamb30	KORE50
Source	Web	Wikipedia	Manual
Long-tail entities	✓	-	✓
# Mentions	15K	1.4M	148

Table 5: ShadowLink in comparison with other datasets that specifically focus on the entity disambiguation task.

on the long-tail entities since the existing benchmarks are known to be biased towards the head of the distribution, i.e., the popular entities (Ilievski et al., 2018; Guo and Barbosa, 2018).

Similarly to ShadowLink, WikiDisamb30 (Ferragina and Scaiella, 2012) contains short text snippets annotated with Wikipedia entities designed for ED evaluation. In contrast to WikiDisamb30, the text snippets in ShadowLink were extracted from web pages outside of Wikipedia to avoid the effects of overfitting since Wikipedia is often used for training language models. Moreover, ShadowLink examples were collected using Wikipedia disambiguation pages as entity spaces while WikiDisamb30 represents a random sample from Wikipedia that does not allow to examine the effect of overshadowing.

The idea of entity spaces was previously introduced by van Erp and Groth (2020), who showed that predicting entity spaces largely improves recall. Their results also hint on the conclusion that disambiguation within entity spaces constitutes a bottleneck in the ED performance. We take this idea further by designing a dataset centered around entity spaces to evaluate ED performance within entity spaces directly. This dataset allows us to measure the gap the state-of-the-art ED systems still have on this task.

KORE50 (Hoffart et al., 2012) was created to evaluate the impact of low commonness and high ambiguity on the ED performance but it contains only 50 hand-crafted sentences with 148 entity mentions including ambiguous mentions and long-tail entities. ShadowLink continues this line of work, providing a considerably larger number of samples that can be used for training and evaluation of ED approaches. We also introduce a subset of neutral samples designed to uncover the model priors. Table 5 summarizes how ShadowLink differs from the previously introduced datasets for entity disambiguation.

Robustness evaluation. Our approach to ED evaluation taps into the fast-growing area of research aimed at assessing model robustness especially relevant for data-driven machine learning techniques. One of the first studies on this topic (Sturm, 2014) argued that the state-of-the art music information retrieval systems show very good performance on the standard benchmarks without the real understanding of the task at hand since their predictions relied solely on the confounds present in the ground truth. Sturm (2014) also coined the term for this phenomena: the "Clever Hans" effect, named after the infamous horse that appeared to solve arithmetic problems while only following unintentional body language cues given by the trainer. More recently, Lapuschkin et al. (2019) showed that the same effect

is demonstrated by other state-of-the-art machine learning models, and the standard performance evaluation metrics fail to detect it. Kauffmann et al. (2020) further explored this phenomenon, showing that it also affects the reliability of unsupervised models in the field of anomaly detection. Therefore, not surprisingly we also observed this effect in the ED task: Guo and Barbosa (2018) used a rudimentary system that merely learned the prior distribution of entities to disambiguate them, and demonstrated that it performs on par with state-of-the-art approaches. These findings specifically calls for new datasets that allow for a more robust evaluation and deeper analysis of the model performance, similar to the one demonstrated here with ShadowLink. We hope that this paper might inspire similar datasets in other fields, where the priors from large public datasets may also overshadow the local context.

6 Conclusion

We introduced ShadowLink, a new benchmark dataset for evaluating entity disambiguation performance, and used it for an extensive analysis of the state-of-the-art systems' results. Our experimental results indicate that all systems under evaluation are prone to rely on their priors, which explains their higher performance on more common entities, and much lower performance on the lexically similar overshadowed entities. Our work thereby shows that the ED task is still far from solved for overshadowed entities, and ShadowLink paves the way for further research in this direction.

The shortcomings of existing disambiguation approaches uncovered by the ShadowLink dataset stimulate further research towards developing more robust ED algorithms that are better at exploiting context without overrelying on the prior bias. We would also like to explore ways to account for more context around the entity mentions, and when expanding the context is actually needed.

Acknowledgements

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the Informatics Institute of the University of Amsterdam, the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the Google Faculty Research Awards program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102.

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, Prague, Czech Republic. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, pages 288–310.

Marieke van Erp and Paul Groth. 2020. Towards entity spaces. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2129–2137, Marseille, France. European Language Resources Association.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1625–1628.

Paolo Ferragina and Ugo Scaiella. 2012. Fast and accurate annotation of short texts with Wikipedia pages. *IEEE software*, 29(1):70–75.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial intelligence*, 194:130–150.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *21st ACM International Conference on Information and Knowledge Management,*

CIKM'12, Maui, HI, USA, October 29 - November 02, 2012, 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.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: an entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2197–2200.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Kauffmann, Lukas Ruff, Grégoire Montavon, and Klaus-Robert Müller. 2020. The Clever Hans effect in anomaly detection. *arXiv preprint arXiv:2006.10609*.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, pages 1–8.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512, Online. Association for Computational Linguistics.

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.

Italo L. Oliveira, Renato Fileto, René Speck, Luís P.F. Garcia, Diego Moussallem, and Jens Lehmann. 2021. Towards holistic entity linking: Survey and directions. *Information Systems*, 95:101624.

Francesco Piccinno and Paolo Ferragina. 2014. From TagME to WAT: a new entity annotator. In *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*, pages 55–62.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625.

René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Named entity recognition using fox. In *International Semantic Web Conference (Posters & Demos)*, pages 85–88. Citeseer.

Bob L Sturm. 2014. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. AGDISTIS - Agnostic disambiguation of named entities using linked open data. In *ECAI*, volume 2014, pages 1113–1114. Citeseer.

Mohamed Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. AIDA: An online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4:1450–1453.