Learn to Not Link: Exploring NIL Prediction in Entity Linking

Fangwei Zhu[1,2], Jifan Yu[1,2], Hailong Jin[1,2], Juanzi Li[1,2], Lei Hou[1,2*] and Zhifang Sui[3]
[1]Dept. of Computer Sci.&Tech., BNRist, Tsinghua University, Beijing 100084, China
[2]KIRC, Institute for Artificial Intelligence, Tsinghua University
[3]MOE Key Lab of Computational Linguistics, Peking University
{zfw19@mails.,yujf18@mails,jinhl15@mails.,lijuanzi@,houlei@}tsinghua.edu.cn
{szf@}pku.edu.cn

Abstract

Entity linking models have achieved significant success via utilizing pretrained language models to capture semantic features. However, the NIL prediction problem, which aims to identify mentions without a corresponding entity in the knowledge base, has received insufficient attention. We categorize mentions linking to NIL into Missing Entity and Non-Entity Phrase, and propose an entity linking dataset NEL that focuses on the NIL prediction problem. NEL takes ambiguous entities as seeds, collects relevant mention context in the Wikipedia corpus, and ensures the presence of mentions linking to NIL by human annotation and entity masking. We conduct a series of experiments with the widely used bi-encoder and cross-encoder entity linking models, results show that both types of NIL mentions in training data have a significant influence on the accuracy of NIL prediction. Our code and dataset can be accessed at https://github.com/solitaryzero/NIL_EL.

1 Introduction

Entity Linking (EL) aims to map entity mentions in free texts to their corresponding entities in a given Knowledge Base (KB). Entity linking acts as a bridge between unstructured text and structured knowledge, and benefits various downstream tasks like question answering (Luo et al., 2018) and knowledge extraction (Chen et al., 2021).

However, not all entity mentions correspond to a specific entity in the KB that suits the mention context (Ling et al., 2015). Take Table 1 as an example, **Peter Blackburn** is actually a journalist, and **the householder** is a common phrase rather than an entity. These two mentions do not refer to any entity in the given KB. The identification of these mentions is referred to as NIL prediction. Therefore, to tackle NIL prediction, the entity linking model needs to select mentions whose references are absent in KB, and link them to a special placeholder *NIL*. Dredze et al. (2010) states NIL prediction as one of the key issues in entity linking, which may lead to a decrease in the recall of entity linking systems. Meanwhile, the incorrectly linked entities may provide false information to downstream tasks.

There have been some earlier representation learning based researches that take NIL prediction into consideration (Eshel et al., 2017; Lazic et al., 2015; Peters et al., 2019). They identify NIL mentions by setting a vector similarity threshold or viewing NIL as a special entity. Recently, pretrained language model (PLM) based models (Wu et al., 2020; Fitzgerald et al., 2021; Cao et al., 2021) have achieved great success for their great transferability and expandability. However, these models generally assume that there always exists a correct entity for each mention in the knowledge base, which leaves the NIL prediction problem without adequate attention.

Previous entity linking datasets have paid insufficient attention to the NIL prediction problem. For example, some of the previous datasets like AIDA (Hoffart et al., 2011) view it as an auxiliary task, while others like MSNBC (Cucerzan, 2007) and WNED-WIKI (Eshel et al., 2017) does not require NIL prediction at all. There does not yet exist a strong benchmark for the ability on NIL prediction.

In this paper, we propose an entity linking dataset NEL that focuses on the NIL prediction problem. About 30% of the mentions in NEL do not have their corresponding entity in the candidate entity set, which requires models to identify these mentions rather than linking them to the wrong candidates. In NEL construction, we take ambiguous entities as seeds, and build the dataset by mining mention contexts related to seed entities on the Wikipedia corpus. Then, human annotators are

[*] Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2023, pages 10846–10860
July 9-14, 2023 ©2023 Association for Computational Linguistics

Missing Entity	
Mention Context	EU rejects German call to boycott British lamb. Peter Blackburn BRUSSELS 1996-08-22
Peter Blackburn (Bishop)	Peter Blackburn (d.1616) was a Scottish scholar and prelate. He was the second Protestant Bishop of Aberdeen.
Peter Blackburn (MP)	Peter Blackburn (1811 – 20 May 1870) was a British Conservative Party politician.
Peter Blackburn (Badminton)	Peter Grant Blackburn (born 25 March 1968) is an Australian badminton player who affiliated with the Ballarat Badminton Association.
Non-Entity Phrase	
Mention Context	Most Hindus accept that there is a duty to have a family during the householder stage of life, as debt to family lineage called Pitra Rin (Father's Debt) and so are unlikely to avoid having children altogether . . .
The Householder (Film)	The Householder (Hindi title: Gharbar) is a 1963 film by Merchant Ivory Productions, with a screenplay by Ruth Prawer Jhabvala . . .
The Householder (Novel)	The Householder is a 1960 English-language novel by Ruth Prawer Jhabvala . . .

Table 1: Example of mentions that should be linked to NIL and their potential candidate entities. Mentions are labeled as red.

asked to identify whether the mentions correspond to a candidate entity or not, and we further perform entity masking to ensure a fair proportion of NIL data of about 30%.

In NIL prediction, we propose to use the widely used bi-encoder and cross-encoder structures as the model backbone, and further integrate type information by adding an entity typing subtask. We combine semantic and type similarity as the final similarity score, and identify NIL mentions by setting a similarity threshold.

We conduct a series of experiments on both NEL and previous entity linking datasets. Experimental results show that the models suffer from an accuracy drop when taking NIL prediction into consideration, indicating that the accuracy may be inflated without the NIL prediction task, and NEL could better diagnose the performance of different models. We also conducted ablation studies on how type information and NIL examples affect the models. We discover that the entity typing subtask yields better embedding even when type similarity is not used, and both types of NIL examples in training data would boost the ability of NIL prediction.

Our contributions can be concluded as:

- We categorize the NIL prediction problem into two patterns: Missing Entity and Non-Entity Phrase, where the latter one has not received sufficient attention in previous works.
- We propose an entity linking dataset NEL focusing on NIL prediction, which covers two patterns of NIL data and could act as a benchmark for diagnosing the ability of NIL prediction.
- We conducted a series of experiments, whose results demonstrate that the accuracy of models may be inflated when not taking NIL prediction into consideration. Meanwhile, both patterns of NIL data in training are essential for triggering the ability of NIL prediction.

2 Preliminary

Entity mentions $M = \{m_i\}$ refer to text spans potentially corresponding to entities in a document $D = (w_1, w_2, ..., w_n)$, where w_i is either a plain token or a mention. Each mention m_i may correspond to an entity $e_i \in E$ in the entity set E of knowledge base $\mathcal{KB}$.

Definition 1. ***Entity linking*** *aims to find an optimal mapping* $\Gamma : M \Rightarrow E$*, which maps entity mentions* $M = \{m_i\}$ *to their corresponding entities* $E = \{e_i\}$*, where* $e_i \in \mathcal{KB}$.

The NIL prediction problem is to determine whether an entity mention m is absent from the knowledge base $\mathcal{KB}$. When there does not exist a proper entity $e \in \mathcal{KB}$ for the given mention m,

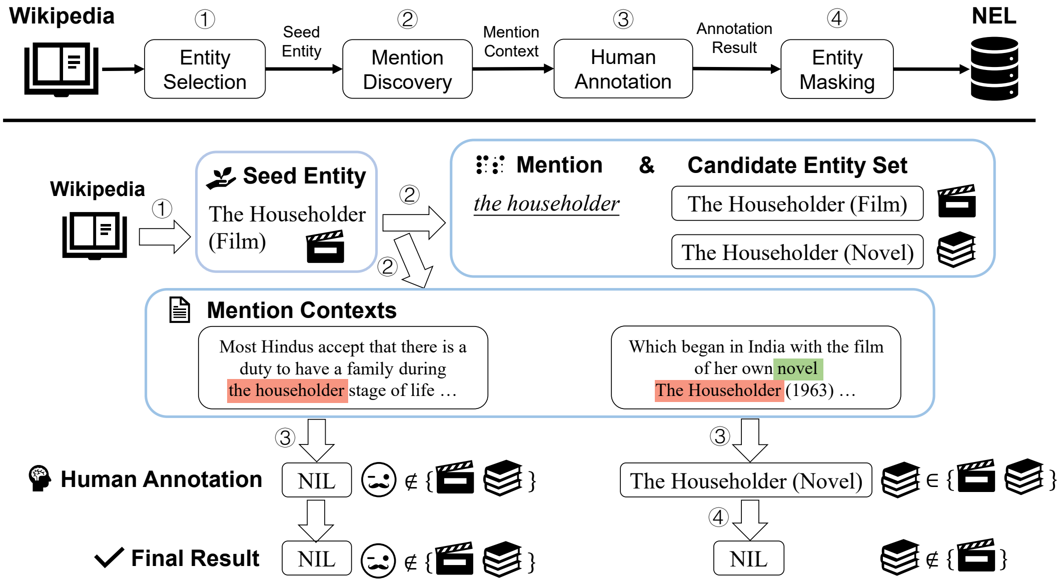

Figure 1: An illustration of how NEL is constructed. We select ambiguous entities as seeds, taking their alias as potential mentions to discover mention contexts from Wikipedia. The entries are annotated by human annotators, and entity masking is performed on some entries to control the portion of NIL data.

the model should link m to a special NIL entity, indicating that the mention is unlinkable.

Definition 2. ***Entity linking with NIL prediction*** *aims to find an optimal mapping $\Gamma : M \Rightarrow E \cup \{NIL\}$, where NIL is a special placeholder and m correspond to NIL only when there is no correct answer in the candidate entity set E.*

As demonstrated in Table 1, there exist two situations in real-world entity linking where NIL prediction should be taken into consideration:

- **Missing Entity:** The mention m refers to certain entity e that has not been yet included in $\mathcal{KB}$, i.e. $m \Rightarrow e \notin \mathcal{KB}$. For example, in the upper half of Table 1, the mention *Peter Blackburn* refers to a certain journalist, while entries in English Wikipedia include only people with other occupations, leading to the mention linking to NIL.

- **Non-Entity Phrase:** The mention m refers to a common phrase that is not usually viewed as an entity, i.e. $m \not\Rightarrow e$. For example, the mention *the householder* in the lower half of Table 1 refers to a concept rather than a film or novel.

3 Dataset Construction

There does not yet exist a strong benchmark on NIL prediction. We manually annotated 300 examples with their mentions linking to NIL from the widely used entity linking dataset AIDA[1], and discover that about 10% of these mentions should actually link to an entity. For example, the mention "EU" in "EU rejects German" should be linked to the European Union rather than NIL (See Appendix D for details). Meanwhile, NIL mentions in AIDA fall mostly in the **Missing Entity** category. The incorrect and imbalanced data for NIL prediction indicates that the importance of NIL prediction is currently underestimated.

In this section, we propose an entity linking dataset named NEL, which focuses on the NIL prediction problem. The construction process is demonstrated in Figure 1.

Unlike normal entity linking data, there does not exist annotated references for mentions linking to NIL, and the portion of NIL data in the text corpus is unknown. Hyperlinks in Wikipedia can be viewed as mentions linking to non-NIL entities, from which we can find the aliases of entities. We assume that when an alias does not appear as a hyperlink in a certain context, it may be identified as a mention linking to NIL. In this way, we collect such contexts as the raw data. The raw data is then annotated by humans to ensure correctness, and we further mask out some answer entities in the candidate set to control the percentage of NIL in

[1]

Dataset	# Data	Annotated	NIL percentage	%Missing Entity	%Non-Entity Phrase
AIDA	34956	✓	20.41%	73%*	10%*
MSNBC	654	✓	0%	-	-
WNED-Wiki	240000	✗	0%	-	-
NEL (ours)	9924	✓	33.57%	17%	83%

Table 2: Statistics of the NEL dataset compared with previous entity linking datasets. *The percentage of two NIL patterns in AIDA is calculated from 300 randomly sampled NIL data, and data with errors do not count as any pattern.

answers.

3.1 Data Collection

Levin (1977) states that the title of creative works could be a place, a personal name, or a certain abstract concept like the choric embodiment of some collectivity (*The Clouds*, *The Birds*) or stock types (*The Alchemist*, *Le Misanthrope*), which would naturally lead to the two situations where a mention links to NIL. The absence of the referenced entity from the KB would lead to **Missing Entity**, while an abstract concept not viewed as an entity would lead to **Non-Entity Phrase**.

To discover NIL mentions of both types, we start by taking entities that share an alias with other entities as seeds. We assume that a mention referring to multiple entities has a higher probability of linking to a **Missing Entity** outside the KB, and the complex meaning of the mention will lead to **Non-Entity Phrase**. Thus, the aliases of ambiguous seed entities would be good starting points for mining NIL mentions.

Entity Selection We further filter ambiguous entities to remove low-quality seeds. First, we remove noise instances like template pages, and entities with less than 5 hyperlinks are also removed. Meanwhile, we discarded entities with a probability greater than 50% of being the linking result, as these entities can generally be considered to be unambiguous and lack difficulty. Finally, 1000 entities are sampled as the seed entity set E_s.

We use a typing system based on Wikidata to identify the type of selected entities. We view the *instance of* relation as the type indicator, and utilize the *subclass of* relation to build a tree-form type system. The detailed type system can be found in Appendix C.

Mention Discovery We build an alias table from the 2021 Wikipedia dump by extracting alias-entity pairs (m, e) from internal hyperlinks. All alias m related to entities in the seed entity set E_s are gathered as the mention set M. For each mention $m \in M$, we look for its occurrence throughout the Wikipedia corpus (whether it appears as a hyperlink or not) to obtain the entry tuple (C_l, m, C_r, E_m), where C_l and C_r represent contexts left and right to the entity mention m, and E_m represents the candidate entities set of m. For each mention m, we sampled 5 entries where m appears as a hyperlink and 5 entries where m appears in plain text to balance the number of positive and negative examples, and a total of 10,000 entries are collected.

3.2 Human Annotation and Post-processing

We perform annotation on the above entries with 3 annotators. The annotators are provided with the mention context (C_l, m, C_r) and candidate entities E_m. Each candidate entity e consists of its title, textual description, and Wikipedia URL. The annotators are asked to choose the answer entity $a \in E_m$ corresponding to the mention m, or $a = NIL$ if none of the candidates are correct.

An expert will further investigate entries in which annotators fail to reach a consensus. The expert is a senior annotator with essential knowledge of entity linking, and will confirm the final annotation result after carefully reading through the context and candidate entities. We use the annotation result as the final answer a if there is an agreement between 3 annotators, and the entity chosen by the expert otherwise. The annotated tuple (C_l, m, C_r, E_m, a) acts as the final entry of our dataset.

To further simulate the situation where new emerging entities do not appear in knowledge bases, we perform entity masking on positive entries. We randomly sample 10% entries where $a \neq NIL$, and mask the correct entity in the candidate set E_m. In this case, as the correct answer is removed from the candidate list, we have $a = NIL$, i.e. the mention m corresponds to the empty entity NIL.

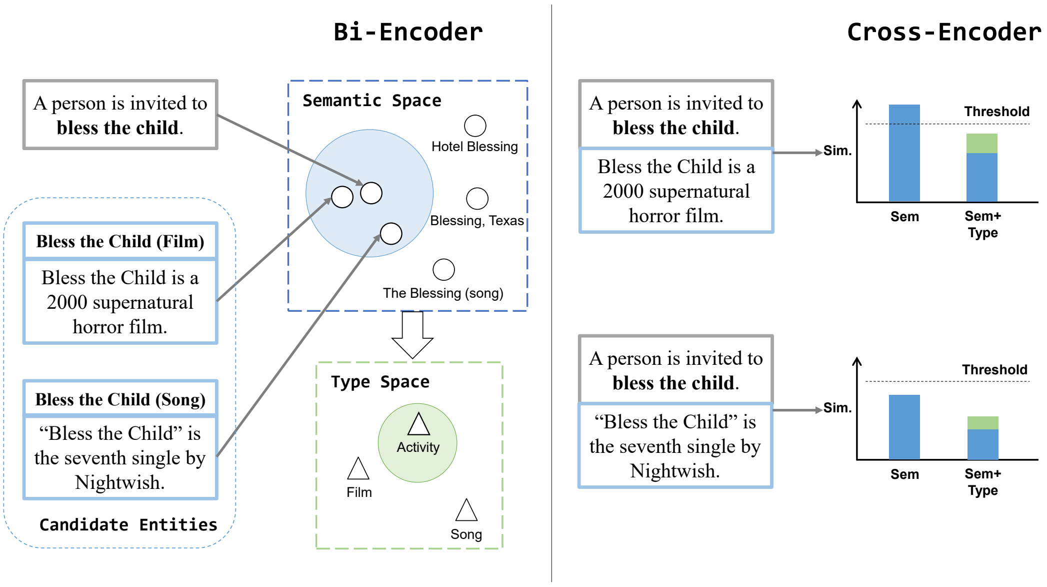

Figure 2: The overall structure of PLM-based retrieval models. Candidates which are confusing in the semantic space may be more distinguishable in the type space. Mentions linking to NIL frequently differ from their candidates in their types, so we combine semantic similarity with type similarity for NIL prediction.

3.3 Dataset Statistics

Table 2 demonstrates the properties of the NEL dataset. NEL includes 6,593 positive examples and 3,331 negative examples, covering 1,000 mentions and 3,840 related entities. Each mention has an average of 3.80 candidate entities. The inter-annotator agreement of NEL is 94.61%, indicating that the expert calibrated about 5% of the data. The full dataset is split into train/validation/test sets by the ratio of 80%/10%/10%.

NEL contains a fair number of entries, and is human-annotated to ensure correctness. Compared with previous entity linking datasets, NEL has a higher percentage of NIL data, which could better diagnose the ability of different models in NIL prediction. Meanwhile, mentions linking to NIL in AIDA mostly fall in the **Missing Entity** situation, while NEL focuses more on the **Non-Entity Phrase** situation, thus complementing the insufficient attention on **Non-Entity Phrase** in NIL prediction.

4 Entity Linking with NIL prediction

A mention links to NIL when all of its candidate entities fail to match its context. A common paradigm of NIL prediction is to compute the similarity scores between the mention context and candidates, and judge its linking result on the base of similarities.

4.1 Scoring Similarities

Bi-encoder and cross-encoder are widely adopted scorer structures, as they are well compatible with pretrained language models. Bi-encoder encodes mention contexts and entities into the same dense vector space, while cross-encoder views similarity scoring as a sequence classification task:

$$s_{bi}(c,e) = \sigma(f(c) \cdot g(e)) \qquad (1)$$

$$s_{cross}(c,e) = \sigma(\mathbf{W}h([c,e]) + \mathbf{b}) \qquad (2)$$

where f, g, h are PLM-based encoders, $\mathbf{W}$ and $\mathbf{b}$ are trainable variables, and σ refers to the sigmoid function.

The bi-encoder structure allows precomputing entity embeddings in knowledge bases, which enables efficient retrieval in real-world applications. Compared with bi-encoder, cross-encoder better captures the relation between context and entities with the cross attention mechanism, thus demonstrating higher accuracy.

4.2 Integrating Entity Types

Previous entity linking models (Gupta et al., 2017; Onoe and Durrett, 2020; Raiman and Raiman, 2018) have proved that entity types do help models better disambiguate between candidate entities.

The type information can be integrated into bi-encoders and cross-encoders by adding a typing layer. In the bi-encoder structure, the mention types

Table 3: Experimental results on NEL and previous datasets. Non-NAC, NAC, and OAC represent non-NIL accuracy, NIL accuracy and overall accuracy. *Results of GENRE on AIDA w/o NIL, MSNBC, and WNED-WIKI are taken from the original paper (Cao et al., 2021).

	NEL (our dataset)			AIDA w/ NIL			AIDA w/o NIL	MSNBC	WNED-WIKI
	Non-NAC	NAC	OAC	Non-NAC	NAC	OAC	OAC	OAC	OAC
BLINK-bi	72.27	88.59	77.74	64.54	**69.59**	65.01	82.61	70.86	58.56
CLINK-bi	79.24	79.28	79.25	75.98	66.36	75.09	83.26	73.29	58.99
GENRE*	54.00	62.84	56.96	-	-	-	**88.60**	88.10	71.70
BLINK-cross	84.09	88.89	85.70	83.08	45.16	79.58	87.49	82.02	69.48
CLINK-cross	**86.97**	**89.19**	**87.71**	**84.42**	58.53	**82.03**	88.16	**89.70**	**72.43**

t_c and entity types t_e are predicted separately:

$$t_c = \sigma(W_c f(c) + b_c) \quad (3)$$

$$t_e = \sigma(W_e g(e) + b_e) \quad (4)$$

while they are simultaneously predicted in the cross-encoder structure:

$$[t_c, t_e] = \sigma(W f([c, e]) + b) \quad (5)$$

where σ represents the sigmoid function and W_c, b_c, W_e, b_e, W, b are trainable parameters.

To tackle the label imbalance between types, we use the focal loss (Lin et al., 2017) on the typing task:

$$\mathcal{L}_t = -\sum_{i=1}^{n_t}(y_i(1-t_i)^\gamma \log t_i + (1-y_i)t_i^\gamma \log(1-t_i)) \quad (6)$$

where n_t is the total number of types in the type system, γ is a hyperparameter, y_i is the golden label of the i-th type and t_i is the predicted label of the i-th type. In bi-encoder, $\mathcal{L}_t$ is the average of loss on t_c and t_e, while in cross-encoder $\mathcal{L}_t$ is directly calculated from $[t_c, t_e]$.

We train the semantic encoder with binary classification loss $\mathcal{L}_s$, and combine $\mathcal{L}_s$ with $\mathcal{L}_t$ as the final loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t \quad (7)$$

4.3 Identifying Mentions Linking to NIL

The type similarity is computed with cosine similarity, and the final score is a weighted sum between type similarity and semantic similarity:

$$s_t(c, e) = cos(t_c, t_e) \quad (8)$$

$$s(c, e) = \lambda s_s(c, e) + (1 - \lambda) s_t(c, e) \quad (9)$$

where λ is a hyperparameter.

For each entry (C_l, m, C_r, E_m, a), we concatenate (C_l, m, C_r) to form the context c. In the training step, for each candidate entity $e \in E_m$, we view (c, e) as a positive example if $e = a$, and as a negative example if $a = NIL$ or $e \neq a$.

During evaluation, the similarity score $s(c, e)$ is computed between context c and each candidate entity e. If there exist entities with a score equal to or higher than the nil threshold $\epsilon = 0.5$, we choose the entity with the highest similarity score as the answer; If all entities fail to reach the threshold, then the mention m links to NIL.

$$a = \begin{cases} \arg\max_e s(c, e), & \exists e, s(c, e) \geq \epsilon \\ NIL, & \forall e, s(c, e) < \epsilon \end{cases} \quad (10)$$

5 Experiments

We conduct a series of experiments on two types of datasets, which test the different ability of entity linking models: (1) NEL that tests the ability of NIL prediction; (2) previous EL datasets that tests the ability of entity disambiguation. We choose the following models for comparison: BLINK (Wu et al., 2020) that uses the bi-encoder and cross-encoder alone to score candidate entities, CLINK that integrates type information with BLINK, and GENRE (Cao et al., 2021) that generates the linking result with a sequence-to-sequence model.

5.1 Main Experiment Results

We trained and tested the models on NEL, with BERT-large as the encoder base of BLINK and CLINK, and BART-large as the backbone of GENRE, to observe their ability in NIL prediction. We also experimented on previous entity linking datasets to observe the disambiguation ability of different models. The models are trained on the AIDA-YAGO2-train (Hoffart et al., 2011) dataset, and tested on AIDA-YAGO2-testb, MSNBC (Cucerzan, 2007) and WNED-WIKI (Eshel et al., 2017).

Table 4: Experimental results on the influence of the entity typing task on NEL. OAC indicates the overall accuracy of entity linking. The overall accuracy with typing is achieved without using the type similarity score.

Model	Ctxt Type Acc.	Cand Type Acc.	OAC w/ Typing	OAC w/o Typing
Bi-encoder	83.75	93.46	78.35	77.74
Cross-encoder	83.95	98.01	87.41	83.67

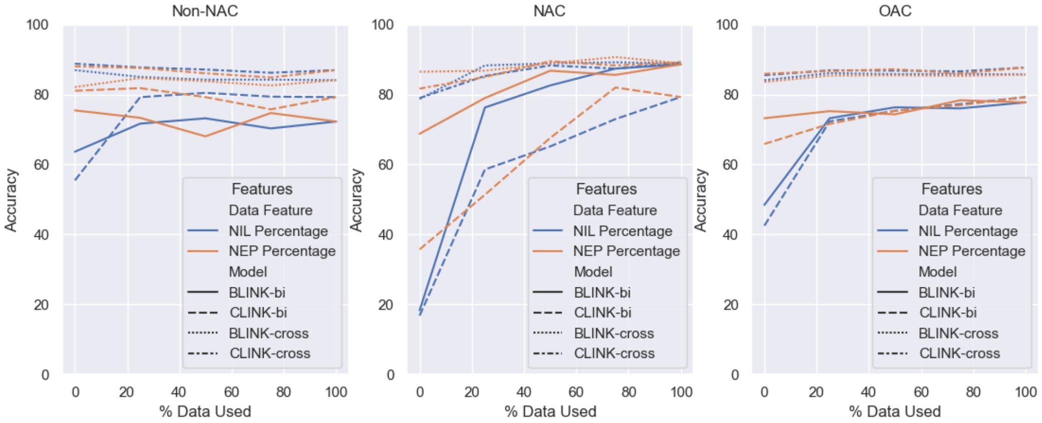

Figure 3: Ablation study on the influence of NIL training data. The x-axis indicates the percentage of used NIL data or Non-Entity Phrase data in the training set.

187 distinct types are used in experiments on NEL, and considering that the entity type distribution may be different across datasets, we use a type system with only 14 top-level entity types on previous datasets to make CLINK more transferable (See Appendix C for details). We retain the same textual representation format with BLINK (see Appendix A), while using 128 as the length of context sequences and entity descriptions. All models are implemented with PyTorch and optimized with the Adam (Kingma and Ba, 2014) optimizer with a learning rate of 1e-5.

Table 3 shows the evaluation results, from which some insights could be discovered:

Type Information Matters. CLINK with cross-encoder structure achieves the highest accuracy on almost all datasets, and is still comparable with GENRE on AIDA, from which we may assert that taking type information into consideration is helpful even without NIL prediction. Meanwhile, on both structures, the overall accuracy of CLINK outperforms BLINK on all datasets, proving that the entity typing task assists both bi-encoder and cross-encoder distinguish correct entities.

Encoder Structure. The cross-encoder structure generally performs better than bi-encoder, but we observe that sometimes bi-encoders show decent ability in detecting NIL mentions, especially when type information is not utilized. BLINK-bi achieves the highest NIL accuracy score of 69.59 on AIDA with NIL, and has a score of 88.59 on NEL, which is comparable with the best-performing CLINK-cross. This phenomenon indicates that cross-encoders may be more prone to overfitting, while entity types would alleviate this tendency.

NIL Entries. On the AIDA dataset, we observe that the models generally suffer from a drop in accuracy when taking NIL entries into consideration, and the drop is more obvious in bi-encoders. This may indicate that the performance of models is inflated without NIL prediction, and NIL entries may confuse the models in practical application.

5.2 Ablation Study

5.2.1 Influence of the Entity Typing Task

We conducted experiments to observe how the entity typing task influences the model. We trained the model with both $\mathcal{L}_t$ and $\mathcal{L}_s$ as loss, while setting $\lambda = 1$ during evaluation to ignore the influence of type similarity scores. Results shown in Table 4 reflects two observations:

First, candidate type predictions benefit from mention context. We observe that when changing the structure from bi-encoder to cross-encoder, the type prediction accuracy on candidates raises by 5%, where the accuracy on contexts remains unchanged. This is likely because the context helps narrow down the type range, while the context type prediction generally remains unaffected by entity descriptions.

Second, unifying the entity typing task in the training process leads to higher overall accuracy, even when the type similarity score is not taken into consideration in the evaluation, which can be demonstrated by the improved score of OAC w/ Typing compared to OAC w/o Typing on both models. This may indicate that predicting entity types would help the model learn semantic embeddings with higher quality.

5.2.2 Influence of NIL Training Data

Compared with previous datasets like AIDA, NEL contains more NIL mentions of the **Non-Entity Phrase** type. We trained the models with different numbers of **Non-Entity Phrase** examples to observe the influence of NIL Training Data.

As demonstrated by Figure 3, all models suffer from a great decline in NIL accuracy when no NIL examples are used during the training stage, and bi-encoder is more prone to the accuracy drop. However, by using only 25% of **Non-Entity Phrase** examples in training, the NIL accuracy would recover to a decent level. Further adding NIL examples has little impact on cross-encoder models, but bi-encoder models still constantly benefit from additional data.

Besides, ignoring NIL data with the **Non-Entity Phrase** type will also harm the NIL accuracy and overall accuracy. Both types of NIL training data are necessary to reach to best performance.

We discover that entity linking models may be unaware of the NIL mentions when there is insufficient training data. A small amount of training data is enough for cross-encoder models to reach a reasonable accuracy, while bi-encoder models constantly benefit from additional training data.

6 Related Work

PLM-based Models in Entity Linking. Using pretrained language models (PLM) to capture semantic information is widely adopted in recent entity linking models. BLINK (Wu et al., 2020) marks the mention in context with pre-defined special tokens, and takes BERT as the encoder base. Two structures are adopted by BLINK to handle different situations: bi-encoder for fast dense retrieval, and cross-encoder for further disambiguation. MOLEMAN (Fitzgerald et al., 2021) searches for similar mention contexts instead of entities, which better captures the diverse aspects an entity reflects in various contexts. GENRE (Cao et al., 2021) finetunes the sequence-to-sequence model BART, directly generating the unique entity name according to the mention context.

Research on NIL Prediction. The NIL prediction problem has been long viewed as an auxiliary task of entity linking. Some entity linking datasets (AIDA (Hoffart et al., 2011), TAC-KBP series (McNamee and Dang, 2009)) take the NIL prediction problem into consideration, while some (ACE and MSNBC) (Ratinov et al., 2011) omit mentions linking to NIL. Some research has already been conducted on the NIL prediction problem. Lazic et al. (2015) and Peters et al. (2019) set a score threshold to filter reasonable candidates, and mentions with no candidate score above the threshold are linked to NIL. Sil and Florian (2016); Kolitsas et al. (2018) views the NIL placeholder as a special entity, and selecting it as the best match indicates that the mention refers to no entities in the given KB. However, recent entity linking models, which use pretrained language models (PLM) as encoder bases, generally take the in-KB setting, which assumes that each mention has a valid golden entity in the KB (Wu et al., 2020).

Entity Type Assisted Entity Linking. Entity types can effectively assist entity linking and have been studied in various works. Gupta et al. (2017) jointly encodes mention context, entity description, and Freebase types with bidirectional LSTM to maximize the cosine similarity. DeepType (Raiman and Raiman, 2018) predicts the type probability of each token and gathers relevant tokens to predict the entity types, which would help eliminate candidates with incompatible types. Onoe and Durrett (2020) views entity types as a training objective rather than a feature, predicting fine-grained Wikipedia category tags to select the most relevant entity.

7 Conclusion

In this paper, we propose an entity linking dataset NEL that focuses on the NIL prediction problem. We observe that mentions linking to NIL can be categorized into two patterns: **Missing Entity** and

Non-Entity Phrase, but the latter one has not been paid sufficient attention. We propose an entity linking dataset NEL that focuses on NIL prediction. The dataset is built upon the Wikipedia corpus by choosing ambiguous entities as seeds and collecting relevant mention contexts. NEL is human-annotated to ensure correctness, and entity masking is further performed to control the percentage of NIL.

We conducted a series of experiments to examine the performance of PLM-based models on different datasets. Experimental results indicate that the accuracy without considering NIL prediction would be inflated. Meanwhile, sufficient data of both NIL types during training is essential to trigger the ability of NIL prediction. In the future, we may further try to integrate entity types into the pretraining process and explore type transfer between datasets.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This paper is supported by a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003).

Limitations

Our work still exist some limitations. First, we choose an entity typing system on the base of Wikidata tags, however, the granularity of the typing system remains to be discussed. A system with too many types would introduce noise to long-tail types, while insufficient types would weaken the disambiguation ability of type similarity. Thus, building a type system with adequate granularity remains a challenge.

Second, we combine the entity typing task with PLM-based semantic encoders, which require a fixed type system and further finetuning. Integrating the entity typing task into the pretraining process may enhance the transferability of the model and remove the dependency on a fixed type system.

Potential Risks. Our proposed dataset NEL centers on ambiguous entities, whose type distribution may not remain the same with other datasets. A potential risk is that the model trained on NEL may experience under-exposure of other entity types, which would damage their transferability and lead to undesired outputs on other datasets.

Ethics Statement

In this section, we will discuss the ethical considerations of our work.

Licenses and terms. The Wikipedia corpus and Wikidata types are obtained via the Wikimedia dump[2], under the CC BY-SA 3.0 license[3]. AIDA, MSNBC, and WNED-WIKI are shared under the CC BY-SA 3.0 license. These datasets have been widely used in entity linking research, and we believe that they have been anonymized and desensitized.

Human Annotation. We recruited 3 human annotators without a background of expertise in annotation, and 1 expert annotator with adequate knowledge in entity linking for checking. These annotators are employed by commercial data annotation companies. We have paid these recruited annotators with adequate rewards under the agreed working time and price. The annotators are well informed about how these annotated data will be used and released, which has been recorded in the contract.

Intended use. NEL is an entity linking dataset focusing on the NIL prediction problem. Researchers are intended to use NEL for examining the ability of NIL prediction of newly created entity linking models. AIDA, MSNBC, and WNED-WIKI are intended for entity linking research, which is compatible with our work.

References

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2021. A lightweight neural model for biomedical entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12657–12665.

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.

[2] https://dumps.wikimedia.org

[3] https://creativecommons.org/licenses/by-sa/3.0/

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 277–285.

Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68.

Nicholas Fitzgerald, Dan Bikel, Jan Botha, Dan Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. Moleman: Mention-only linking of entities with a mention annotation network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 278–285.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.

Harry Levin. 1977. The title as a literary genre. *The Modern language review*, 72(4):xxiii–xxxvi.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2185–2194.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text analysis conference (TAC)*, volume 17, pages 111–113.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8576–8583.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2255–2264.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

A Details about the NEL Dataset Construction

A.1 Corpora

The NEL dataset is built from the 2021-07 English Wikipedia dump under the CC BY-SA 3.0 license. We take the hyperlinks in the raw xml dump as entity mentions, and retain at most 128 tokens around the mentions as their context. We take 64 tokens left to the mention and 64 tokens right to the mention by default, and more tokens will be included in one side if the other side does not contain enough

	NEL & AIDA w/ NIL			
	CLINK-bi		CLINK-cross	
Hyperparameter	Value	Range	Value	Range
Learning Rate	1e-5	{1e-3, 1e-5, 3e-5}	1e-5	{1e-3, 1e-5, 3e-5}
λ	0.5		0.5	
ϵ	0.5		0.5	
Epoch	4	{1, 4}	4	{1, 4}
Batch Size	4	{1, 4, 8, 16, 32}	1	{1, 4, 8, 16, 32}
# Parameters	670M	-	335M	-
Training Time	$\tilde{2}$ hrs	-	$\tilde{3}$.5 hrs	-
	AIDA w/o NIL, MSNBC & WNED-WIKI			
	CLINK-bi		CLINK-cross	
Hyperparameter	Value	Range	Value	Range
Learning Rate	1e-5	{1e-3, 1e-5, 3e-5}	1e-5	{1e-3, 1e-5, 3e-5}
λ	0.9	[0.0, 1.0]	0.9	[0.0, 1.0]
ϵ	0	-	0	-
Epoch	1	{1, 4}	1	{1, 4}
Batch Size	4	{1, 4, 8, 16, 32}	1	{1, 4, 8, 16, 32}
# Parameters	670M	-	335M	-
Training Time	$\tilde{4}$ hrs	-	$\tilde{3}$ hrs	-

Table 5: The hyperparameters used in the training process.

tokens. We then discard tokens from both ends to ensure that the context plus the mention do not exceed the 128 token limit. Media files (image, audio) and Lua commands are discarded during preprocess.

A.2 Data Selection

Entries with the following features are viewed as noise and discarded:

- The mention context contains the token '*', which is usually a list or formula;
- The mention with only 1 candidate entity, which does not pose much challenge;
- The mention with more than 20 candidate entities, which is far too challenging;
- The mention appears as a sub-span of a word;
- The mention has a probability of over 50% of linking to a certain candidate entity, in which case we view the mention as unambiguous.

A.3 Textual Representation Format

Textual representation format for bi-encoder:

$$C = \text{[CLS]}\ C_l\ [m_{start}]\ m\ [m_{end}]\ C_r\ \text{[SEP]}$$
$$E = \text{[CLS]}\ e_{title}\ [m_{title}]\ e_{desc}\ \text{[SEP]}$$

Textual representation format for cross-encoder:

$$(C, E) = \text{[CLS]}\ C_l\ [m_{start}]\ m\ [m_{end}]\ C_r\ \text{[SEP]}\ e_{title}\ [m_{title}]\ e_{desc}\ \text{[SEP]}$$

where C represents mention context and E represents the textual description of candidate mentions. $[m_{start}]$, $[m_{end}]$, $[m_{title}]$ are special tokens.

B Experiment Details

We use the BERT-large-uncased model as the encoder base, with parameters initialized from the python *transformers* library. The models are trained on a single NVIDIA GeForce RTX 3090 GPU. We obtain the AIDA, MSNBC and WNED-WIKI dataset from the BLINK (Wu et al., 2020) repository https://github.com/facebookresearch/BLINK. We trained our model on the AIDA-train split, and evaluated on all three datasets.

The hyperparameter configurations are as follows. Detailed hyperparameters are shown in Table 5.

C Typing System

C.1 Typing System on NEL

We use a tree-like typing system with 187 distinct types on NEL. The typing system is build on the

Table 6: Examples of type lines

Entity	Types
14th Street (Manhattan)	Road->RouteOfTransportation->Infrastructure->ArchitecturalStructure->Place
1958 Copa del Generalísimo	SoccerTournament->Tournament->SportsEvent->SocietalEvent->Event
ATM (song)	Song->MusicalWork->Work
Brats (1991 film)	Film->Work
Babe Ruth	BaseballPlayer->Athlete->Person

base of Wikidata types. Table 6 shows some examples of type lines in the system. The most 10 frequent types in NEL are: (Work, Organisation, Place, Event, Person, Activity, FictionalCharacter, Award, Species, MeanOfTransportation).

C.2 Typing System on Traditional Datasets

We retain 14 top-level types to make CLINK more transferable on different datasets. These types are: (Other, Person, Place, Work, Organization, Event, Fictional Character, Species, Activity, Device, Topical Concept, Ethnic Group, Food, Disease)

D Errors in AIDA

Table 7 demonstrates some mentions in AIDA that are incorrectly linked to NIL. 50 errors are detected among the 300 randomly sampled data in AIDA.

E Case Study

Table 8 shows some examples predicted by CLINK and BLINK without type information, which reflects how entity types influence the linking result.

In the first example, models with the bi-encoder structure incorrectly take the "Gates of Heaven" entry (which is in fact a documentary film) as the linking result, while CLINK-cross notices the context word "album" may indicate the entity type type, and correctly links the mention to the album. In the second example, the "Home Before Dark" mention actually refer to a 1997 movie[4] like other movies in the context, however the corresponding entry is absent in the English Wikipedia. The CLINK-cross model is able to identify that the mention should be labelled as NIL, where the other models mistakenly link it to the "Home Before Dark" entry, which is an album rather than a movie. We observe that the entity types do help measure the similarity between mentions and entities, which enhances the performance of CLINK.

[4] https://www.imdb.com/title/tt0116547/

Mention Context	Mention	Assumed Entity
Bosnian premier in Turkey for one day visit . ANKARA 1996-08-27	Turkey	Turkey (Country)
EU rejects German call to boycott British lamb . Peter Blackburn BRUSSELS 1996-08-22	EU	European Union
U.S. F-14 catches fire while landing in Israel . JERUSALEM 1996-08-25 A U.S. fighter plane blew a tyre and . . .	F-14	Grumman F-14 Tomcat
This is the leading story in the Mozambican press on Monday. Reuters has not verified this story and does not vouch for its accuracy.	Reuters	Reuters

Table 7: Examples data in AIDA that are incorrectly linked to NIL.

Mention Context:	. . . singer, Michelle Branch, during her visit in Japan to promote her album Hotel Paper, for a magazine interview and photoshoot. After the release of Gates of Heaven the group a short break and performed in New York City . . .			
Model	BLINK-bi	CLINK-bi	BLINK-cross	CLINK-cross
Prediction	Gates of Heaven	Gates of Heaven	NIL	**Gates of Heaven (album)**
Mention Context:	. . . Michael Williams and David Collins founded the company in 1994 focusing on independent features, including Never Met Picasso (1996), Home Before Dark (1997), Six Ways To Sunday (1998), . . .			
Model	BLINK-bi	CLINK-bi	BLINK-cross	CLINK-cross
Prediction	Home Before Dark	Home Before Dark	Home Before Dark	**NIL**

Table 8: Examples of predicted entity by different models. Entity mentions in context are labelled as red and the correct answer is labelled as **bold**.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

□ B1. Did you cite the creators of artifacts you used?
No response.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section Ethical Considerations

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section Ethical Considerations

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section Ethical Considerations

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section Ethical Considerations

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The instruction materials contain private information about authors and annotators, which may affect the double-blind review process. We have not yet been permitted to publicate it.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section Ethical Considerations

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section Ethical Considerations

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
There have not yet established an ethics review board committee in our region.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
These data involve private information about annotators, and we have not yet been permitted to publicate it.