Hansel: A Chinese Few-Shot and Zero-Shot Entity Linking Benchmark

Anonymous ACL submission

Abstract

Modern Entity Linking (EL) systems entrench a popularity bias. However, there is no dataset focusing on tail and emerging entities in languages other than English. We present Hansel, a new benchmark in Chinese that fills the vacancy of non-English few-shot and zero-shot EL challenges. The test set of Hansel is human annotated and reviewed, created with a novel method for collecting zero-shot EL datasets. It covers 10K diverse documents in news, social media posts and other web articles, with Wikidata as its target Knowledge Base. We demonstrate that the existing state-of-the-art EL system performs poorly on Hansel (R@1 of 36.6% on Few-Shot). We then establish a strong baseline that scores a R@1 of 46.2% on Few-Shot and 76.6% on Zero-Shot on our dataset. We also show that our baseline achieves competitive results on TAC-KBP2015 Chinese Entity Linking task.

1 Introduction

Entity Linking (EL) is the task of grounding a textual mention in context to a corresponding entity in a Knowledge Base (KB). It is a fundamental component in applications such as Question Answering (Févry et al., 2020a; Guu et al., 2020; De Cao et al., 2019), KB Completion (Shen et al., 2014; Zhang et al., 2014) and Dialogue (Curry et al., 2018).

An unresolved challenge in EL is to accurately link against emerging and less popular entities. The Zero-Shot Entity Linking problem was presented by Logeswaran et al. (2019), aiming at linking mentions to entities unseen during training. On the other hand, Chen et al. (2021) raised a common popularity bias in EL, i.e. EL systems significantly under-perform on tail entities that share names with popular entities. Intuitively, we name the challenge to resolve tail entities as *Few-Shot Entity Linking*, as most of them have only a few number of training examples. Despite the aforementioned studies, non-English resources for zero-shot and few-shot EL are seldom available, hindering progress for these challenges across languages.

Moreover, existing zero-shot and few-shot EL datasets have a limited diversity, rooted from their collection methods that rely on hyperlink structures or manual templates. Logeswaran et al. (2019) extracted mentions from Wikia articles hyperlinked to the Wikia KB, and Botha et al. (2020) used links from Wikinews to Wikipedia, where only 3K out of 289K (1%) mentions fall into its zero-shot slice. Chen et al. (2021) generated AmbER sets by filling pre-defined templates with KB attributes. These dataset collection approaches are limited, as mentions are biased towards hyperlink editing conventions or syntactic templates.

To address the language bias and lack of syntactic diversity in few-shot and zero-shot EL datasets, we present Hansel, a human-calibrated and challenging EL benchmark in simplified Chinese. Hansel consists of few-shot and zero-shot test sets, as well as a Wikipedia-based training set. The few-shot slice is collected from a multi-stage matching and annotation process. A core property of this dataset is that all mentions are ambiguous and "hard" (Tsai and Roth, 2016), where the ground-truth entity is not the most popular by the mention. The zero-shot slice is collected from a novel searching-based process, where annotators are presented with a new entity's description, and find corresponding mentions and adversarial examples with Web search engines over diverse domains. We demonstrate that both slices are challenging for state-of-the-art EL models. We further design a type system exploiting rich Wikidata structure, and propose a novel architecture utilizing the type system that improves over dual-encoder based models.

The main contributions of this work are:

- Publish Hansel, a challenging multi-domain benchmark for Chinese EL with Wikidata as KB, featuring a zero-shot slice with emerging entities, a few-shot slice with hard mentions, and a large training set with 1M documents.

- Propose a novel and feasible zero-shot entity linking dataset collection paradigm, applicable for any language.
- Achieve strong results on TAC-KBP2015 Chinese EL task with a monolingual model, on a par with state-of-the-art multilingual models on this task.

2 Related Work

For years, the primary focus of Entity Linking studies were constrained to English-only and fixed-KB settings (Ling et al., 2015; Févry et al., 2020b; Ling et al., 2020; De Cao et al., 2021a). Cross-Lingual Entity Linking (XEL) was introduced to link non-English mentions to English KBs. (McNamee et al., 2011; Ji et al., 2015) Recently, Botha et al. (2020) introduced **Multilingual Entity Linking**, a more general formulation to link mentions from any language to a language-agnostic KB. Their Mewsli-9 multilingual benchmark alleviates the language bias in general EL to some extent, but many languages including Chinese are not yet covered.

Zero-Shot Entity Linking was proposed by Logeswaran et al. (2019), with an English zero-shot EL dataset published. Mewsli-9 has a zero-shot slice of 3,198 multilingual mentions, though only hyperlinked texts in Wikinews are included. Zero-shot EL on temporally evolving KBs has been less discussed. To this end, Hoffart et al. (2014) proposed EL on emerging entities, but the dataset is also English-only. In this work, we present the first non-English zero-shot EL dataset focusing on emerging entities.

Few-Shot Entity Linking was frequently studied recently. Provatorova et al. (2021) suggested that it is possible to obtain high accuracy on popular EL datasets by merely learning the prior, and released ShadowLink test set whose "Shadow" subset is similar with our few-shot setting, but only available in English. Chen et al. (2021) discovered that current EL systems significantly under-perform on tail entities, and released AmbER test sets for this task. Their dataset is English-only and generated by filling pre-defined templates with KB attributes. Tsai and Roth (2016) has a cross-lingual "hard" subset similar to our setting, but the corpus domain is limited to Wikipedia. In this work, we present the first non-English, human-calibrated few-shot EL dataset with better syntactic diversity.

In Chinese language, existing EL datasets are very limited. An established dataset is TAC-KBP2015 Tri-Lingual Entity Linking Track (Ji et al., 2015), adapting the Cross-Lingual EL setting where the mention is in Chinese and the KB is in English. Datasets in the same series (Ji et al., 2016, 2017) are also relevant. DuEL (Han et al., 2020) is an EL dataset with a native Chinese KB, but the KB only includes an incomplete subset of Baidu's knowledge base (390K entities), making it difficult to serve as a comprehensive EL benchmark. CLEEK (Zeng et al., 2020) contains 2,786 mentions, annotated to the union of Chinese Wikipedia and CN-DBPedia (Xu et al., 2017), but it does not focus on zero-shot or few-shot EL. More comparison of existing Chinese EL benchmarks and their limitations are in Appendix I. Our proposed benchmark enriches Chinese EL resources and alleviates their popularity bias, providing basis for Chinese and multilingual few-shot and zero-shot EL studies.

3 Hansel Dataset

Define a set of **entities** E that are entries of a Knowledge Base (KB). Given an input text document $D = \{s_1, \ldots, s_d\}$ and a set of **entity mentions** that are spans with known boundaries: $M_D = \{m_1, \ldots, m_n\}$, an Entity Linking (EL, also referred to as Entity Disambiguation) system outputs mention-entity pairs: $\{(m_i, e_i)\}_{i\in[1,n]}$, where each entity is either a known KB entity or NIL (an entity out of KB): $e \in E \cup \{nil\}$. The broader instance of EL where mention spans are not given De Cao et al. (2021a) is out of scope for this work.

We publish an EL dataset for simplified Chinese (zh-hans), named Hansel. The training set is processed from Wikipedia. The test set of Hansel contains Few-Shot (FS) and Zero-Shot (ZS) slices, focusing respectively on tail entity linking and zero-shot generalization to emerging entities. Both test sets contain mentions drawn from diverse documents, with the ground truth entity ID annotated. Dataset statistics are shown in Table 1.

3.1 Knowledge Base

To reflect the common scenario of temporally evolving knowledge bases, we split Wikidata entities into **Known** and **New** sets using two historical dumps:

Known Entities (E_{known}) refer to Wikidata entities in 2018-08-13 dump. All our models are trained with E_{known} as KB.

New Entities (E_{new}) refer to Wikidata entities

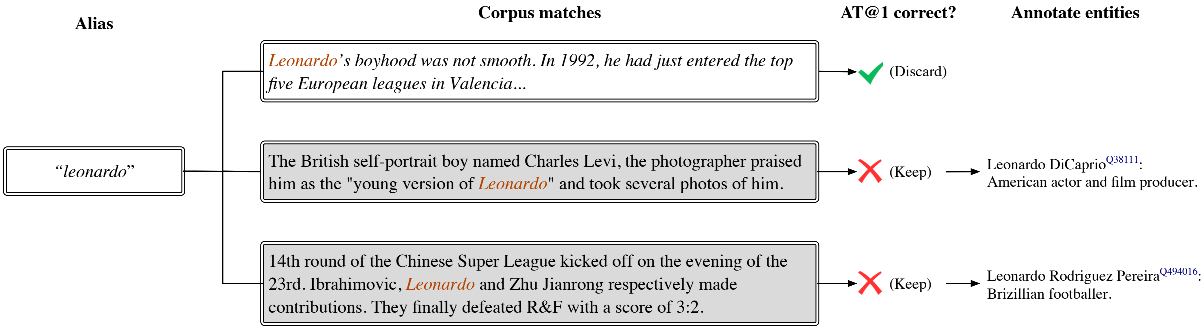

Figure 1: **Annotation process for the Few-Shot dataset**, with an actual (translated) example in Hansel-FS. We first match aliases against the corpora to generate diversified potential mentions, then annotate if the most popular entity (AT@1) is the correct candidate for each mention. We only keep cases where AT@1 is incorrect, and annotate the correct entity against the KB.

Figure 2: **Annotation process for the Zero-Shot dataset**, with a translated example in Hansel-ZS. Given a new entity, we search on the Web for a corresponding mention, and a few mentions that share the same mention text but refer to different entities.

in 2021-03-15 dump that do not exist in E_{known}. Intuitively, entities in E_{new} were newly added to Wikidata between 2018 and 2021 thus never seen when training on 2018 data, thus considered as a zero-shot setting.

Entity filtering. We filter original Wikidata entities extending logic by Botha et al. (2020) to get a clean KB: we remove all instances of Wikimedia disambiguation pages, templates, categories, modules, list pages, project pages, Wikidata properties, as well as their subclasses, as detailed in Appendix D. For the scope of this paper, we further constrain to entities with a Chinese Wikipedia page (in Wikipedia 2021-03-01 dump). After filtering, E_{known} contains roughly 1M entities and E_{new} contains 57K entities.

Alias table. An alias table defines the probability of a text mention m linking to an entity e, i.e. $P(e|m)$. We extract an alias table from Wikipedia 2021-03-01 for both E_{known} and E_{new} by parsing Wikipedia internal links, redirections and page titles, following (De Cao et al., 2021b). We denote this alias table as *AT-base*.

Wikidata Type system. Prior work demonstrated that types can benefit EL systems (Ling et al., 2015; Raiman and Raiman, 2018; Fu et al.). We introduce a new formulation for coarse and fine entity typing, utilizing rich structural knowledge in Wikidata. The type system is general and language-agnostic. Define original Wikidata entities as E, property types as P, and relation triples as $R(e_1, p, e_2)$. We define a transitive typing feature denoted as $Type$:

$$R(e_1, P31, e_2) \Rightarrow Type(e_1, e_2),$$

$$Type(e_1, e_2) \wedge R(e_2, P279, e_3) \Rightarrow Type(e_1, e_3),$$

where P31 stands for *instance of* and P279 for *subclass of* relations in Wikidata. We then define coarse types with this feature:

Coarse Types. We define in Table 2 five orthogonal categories: person (PER), location (LOC), organization (ORG), event (EVENT), and others (OTHER). Note that our location type effectively combines GPE, LOC and FAC types as defined in ACE (Doddington et al., 2004) and TAC-KBP2016 (Ji et al., 2016) in order to better fit Wikidata typing guideline [1]. We use the same PER definition as

[1]We refer to https://www.wikidata.org/wiki/Wikidata:WikiProject_Infoboxes when choosing appropriate entities for corresponding types.

	# Mentions			# Documents			# Entities		
	In-KB	NIL	Total	In-KB	NIL	Total	E_{known}	E_{new}	Total
Train	9.89M	-	9.89M	1.05M	-	1.05M	541K	-	541K
Validation	9,677	-	9,677	1,000	-	1,000	6,323	-	6,323
Hansel-FS	3,404	1,856	5,260	3,389	1,850	5,234	2,720	-	2,720
Hansel-ZS	4,208	507	4,715	4,200	507	4,704	1,054	2,992	4,046

Table 1: **Statistics of the Hansel dataset.** We break down the number of mentions and documents by whether the label is a NIL entity or inside Wikidata (In-KB), and the number of distinct entities by whether the entity is in an emerging entity in E_{new}.

Coarse Type	Definition
$PER(e)$	$Type(e, Q215627)$
$LOC(e)$	$Type(e, Q618123)$
$ORG(e)$	$Type(e, Q43229)$
$EVENT(e)$	$Type(e, Q1656682)$
$OTHER(e)$	All other entities

Table 2: Coarse types defined with transitive $Type$.

TAC-KBP2016, and add an EVENT type.

Fine Types. We design an entity feature *TopSnaks* as our fine typing system. TopSnaks are defined as top 10,000 property-value pairs, i.e. (p, e_2) tuples, sorted by entity frequency in KB [2]. An example TopSnak is *P31-Q5*, which means "instance of human". We verify that the TopSnaks generated on the 2018 Wikidata dump covers about 90% of E_{new}, indicating good generalizability over time. Examples of TopSnaks are in Appendix C.

3.2 Training Data

Following previous work (Botha et al., 2020; De Cao et al., 2021a), we use Wikipedia internal links to construct a training set. The alignment of Wikidata and Wikipedia ecosystems enables utility of rich hyperlink structure in Wikipedia.

All new entities E_{new} are kept unseen during training. Ideally, one would acquire the 2018 Wikipedia dump as the training corpus. As the full 2018 Wikipedia dump is not publicly available, we use 2021-03-01 Wikipedia dump and hold out all entity pages mapped to E_{new} as well as all mentions with pagelinks to E_{new} entities. Our zero-shot evaluation slice is based on E_{new}.

To focus on simplified Chinese, we consider Chinese-Wikipedia only, and converted traditional Chinese characters to simplified in all training and evaluation sets, as well as the alias table [3]. The training set contains 9.9M mentions from 1.1M documents. We hold out 1K full documents (9.7K mentions) as the validation set.

3.3 Few-Shot Evaluation Slice

For the Few-Shot (FS) test set, we collect human annotations in three Chinese corpora: LCSTS (Hu et al., 2015), covering Weibo microblogging short text, SohuNews and TenSiteNews, long news articles from Sohu and other news sites (Wang et al., 2008). Details of these corpora are in Appendix A.

Matching. The FS slice is collected based on a matching-based process as illustrated in Figure 1. We first use *AT-base* to match against the corpora to generate candidates, then sample ambiguous mentions diversified by mention-text for human annotation. Note that we only match ambiguous mentions with at least two entity candidates in E_{known}, and keep limited examples per mention for better diversity. Matching and sampling details are in Appendix A.

Annotation. Human annotation was performed on more than 15K examples with 15 annotators. For each example, annotators answer a series of questions: First, they modify the incorrect mention boundary, or remove the example if it is not an entity mention. Then, they select among alias table candidates for the referred entity. For each candidate, annotators have access to its entity description (first paragraph in Wikipedia) and the original Wikipedia link. If the candidate with the highest prior (AT@1) is correct, then the example is discarded. 75% of examples are dropped in this step. If none of the candidates are correct, the annotator is then asked to find the correct Wikipedia page (mapped to a Wikidata QID) for the entity through search engines. If no Wikipedia page can

[2]"SNAK" is a Wikidata term referring to "some notation about knowledge" : https://www.wikidata.org/wiki/Q86719099.

[3]We use HanziConv to convert to Simplified Chinese: https://github.com/berniey/hanziconv.

be found, they fill the coarse entity type defined in Table 2 and label a NIL entity with its coarse type. The process is illustrated in Figure 2. More examples of the FS slice are in Appendix E.

3.4 Zero-shot Evaluation Slice

Collecting a Zero-Shot (ZS) slice is challenging, due to the difficulty to find occurrences of new entities on a fixed text corpus, especially when the corpus is out-of-domain and hyperlink structures cannot be exploited. To address this challenge, we design a novel data collection scheme by searching entity mentions across the Web given an entity description. The process is detailed below.

Type balancing. We first down-sample E_{new} to get a diverse set of entities with various coarse types, as the original distribution of E_{new} is heavily biased towards OTHER (52%) and PER (38%). We draw samples from E_{new} by 50% random sampling and 50% type-diversified sampling.

Searching-based Annotation. For each entity in E_{new}, annotators are given its title, description and Wikidata aliases. They are asked to search the Internet [4] for a corresponding mention of the entity, and collect the mention context. They further seek 1 or 2 adversarial examples by searching for a same or similar mention referring to a different entity. The process is illustrated in Figure 2 with an example of adversarial examples. Such confusing examples introduce more label diversity and reduce bias on this dataset. More examples of the ZS slice are in Appendix E.

3.5 Dataset Quality and Statistics

Expert checking. For both FS and ZS slices, after the first pass of annotation, there is an expert-checking phase, where 5 human experts manually examine and correct all annotated examples. "Experts" are well-trained annotators who made fewest mistakes in the trial annotation and learned basic knowledge of entity linking. Each example is labeled by one annotator and reviewed by one expert (i.e. tie-breaking by choosing the expert's result). The expert-reviewed results are used as the ground truth (GT) of this dataset.

Dataset statistics. As reported in Table 1, the FS slice has 5,260 mentions from 5,234 documents, covering 2,720 diverse entities. The ZS slice has 4,715 mentions across 4,707 documents, covering 4,046 distinct entities. Domains of examples are in news (51.5%) and social media (48.5%) for FS slice, and news (38.6%), social media (14.9%), and other articles such as E-books, papers and commerce (46.4%) for ZS slice.

Dataset Quality. To measure dataset quality, we first calculate the percentage agreement between the annotator and the expert. The percentage agreement of Hansel-FS and Hansel-ZS are 87.3% and 95.9% respectively, i.e. modification rate is 12.7% and 4.1% during expert checking. Both imperfect mention boundaries and wrong entities count as disagreements, whereas boundary changes account for 40.1% for FS disagreements and 53% for ZS.

We further take a random sample from the final dataset, 100 entries from FS and 100 from ZS, and present the mention context with the GT entity to two annotators, to independently label whether GT is correct. In this step, two annotators agree on 88% of the cases in FS slice and 94% of the cases in ZS slice. We use Cohen's Kappa coefficient to evaluate the inter-annotator agreement. The coefficient is 0.622 for FS and 0.651 for ZS, indicative of substantial agreement between annotators (Fleiss and Cohen, 1973). Average human accuracy (evaluating on GT) is 88% for FS and 95.5% for ZS.

3.6 Extending to Other Languages

To port our annotation method to a new language, one may re-use our chosen Wikidata dumps to construct E_{new} and E_{known}, and apply a different language filter to get the target set of entities. Then, one may obtain an alias table by parsing language-specific Wikipedia. If there is a large text corpus for the language, one may adopt our matching-based process in Section 3.3 for a few-shot EL dataset. For new entities (also applicable for few-shot entities, if no matching corpus is available), one may refer to the searching-based method in Section 3.4, to present annotators each entity and search the Web for mentions in the language. The annotators need to have expertise in the target language.

4 Models

We establish baseline models on the Hansel dataset, including a Dual Encoder (DE) model and a Cross-Attention encoder (CA) model for entity disambiguation. We also present a novel architecture that exploit our coarse and fine typing system, and show that typing-based auxiliary supervision provides improvements on DE.

[4] To facilitate searching, we provide annotators with pre-filled search query templates in an annotation tool, such as Google queries with entity names and target domains.

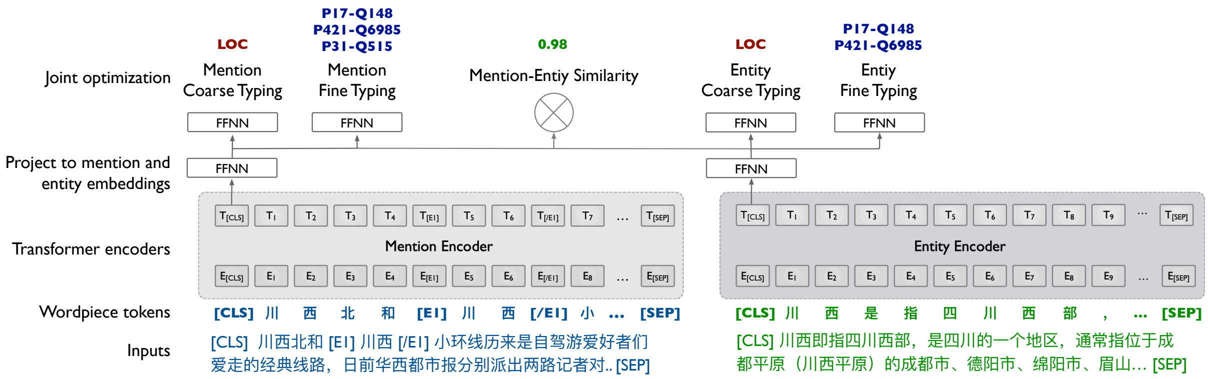

Figure 3: **Typing-enhanced Dual Encoder (TyDE)** architecture. Both mention and entity encoders are 12-layer transformer encoders initialized from BERT-base, projecting mention in context (annotated with [E1] and [/E1] markers) and entity description to 256-d embeddings. Cosine similarity between mention and entity embeddings is jointly optimized with typing losses.

4.1 Dual Encoder Model

Following previous work (Wu et al., 2020; Botha et al., 2020), we train a Dual Encoder (DE) model to project entity and mention contextual representations into a same vector space. Such models are scalable in that the entity embeddings can be pre-computed and stored, enabling fast retrieval or dot-product based similarity scoring.

The dual encoder takes a mention-entity pair (m, e) and outputs their cosine similarity score:

$$sim(m, e) = \frac{\phi(m)^T \psi(e)}{\|\phi(m)\| \|\psi(e)\|}, \quad (1)$$

where both ϕ and ψ are learned transformer encoders projecting mention and entity input sequences into d-dimensional vectors (d=256). For both encoders, we use BERT-base and map the $[CLS]$ token with a dense layer to the output embedding. Following Botha et al. (2020), we use mention boundary tokens to wrap mentions in context. We concatenate the title and the first paragraph in Chinese Wikipedia as an entity's description for input of ψ. The DE model is optimized with in-batch sampled softmax loss.

We use the DE model as a scoring step on candidates generated by the alias table *AT-base*, combining the model's prediction $sim(m, e)$ with the prior $P(e|m)$ to produce a score $s(m, e)$:

$$s(m, e) = P(e|m) sim(m, e). \quad (2)$$

4.2 Cross-Attention Encoder Model

Following Botha et al. (2020), we train a Cross-Attention encoder model (CA) which takes concatenated mention and entity inputs, the same text representations as for DE, and encodes their similarity. We optimize CA with a binary cross-entropy loss. We use CA's output score to rank candidates generated by the alias table.

Since the training set only comes with positive examples, we use the alias table to mine hard negatives, and randomly keep 20% of negative examples to reduce label imbalance.

4.3 TyDE: Typing-enhanced Dual Encoder

Previous work (Ling et al., 2015; Raiman and Raiman, 2018) suggested that type coherence can benefit EL systems. However, models like DE or CA only implicitly learn type coherence with pre-trained contextualized representations. Moreover, types for new entities in KB can be incomplete.

We propose a novel model architecture, typing-enhanced dual encoders (TyDE), using Wikidata type system as an auxiliary supervision task to improve the dual encoder model. On top of mention and entity encodings output by ϕ and ψ, we add classification layers for coarse and fine typing. On each side, we use a softmax classifier for coarse types and binary classifiers for each of the 10K fine types. We train the TyDE model with positives only, using type classification losses in addition to the batch softmax loss, illustrated in Figure 3. The supervision approach does not rely on types as encoder input, thus less prune to KB incompleteness and does not require types for inference.

During inference, we use the similarity score as defined in DE, $P(e|m)sim(m, e)$, and combine it with coarse and fine typing scores. Coarse typing score S_c and fine typing score S_f are defined as:

$$\begin{aligned} s_c(m, e) &= \sigma_c(m)^T \rho_c(e), \\ s_f(m, e) &= \sigma_f(m)^T \rho_f(e) \end{aligned} \quad (3)$$

where σ_c, ρ_c, σ_f and ρ_f are single linear dense

	Metric	Value
Tsai and Roth (2016)	R@1	85.1
Sil et al. (2018)	R@1	85.9
Upadhyay et al. (2018)	R@1	86.0
Zhou et al. (2019)	R@1	85.9
De Cao et al. (2021b)	R@1	**88.4**
DE	R@1	75.2
TyDE	R@1	76.2
CA	R@1	81.7
CA-tuned	R@1	<u>88.1</u>
AT-base	R@1	73.1
AT-base	R@10	89.1
AT-base	R@100	89.4
AT-ext	R@1	75.3
AT-ext	R@10	91.1
AT-ext	R@100	91.5

Table 3: Recall evaluations on the TAC-KBP2015 Chinese EL task. Our monolingual CA-tuned model is on a par with the multi-lingual SOTA. We also report recall with our base and extended alias tables.

layers, projecting ϕ and ψ outputs to corresponding type dimensions. σ_c and ρ_c project to 5 coarse types, and σ_f and ρ_f project to 10,000 fine types.

We experiment TyDE for scoring with different settings: (1) similarity only, i.e. $P(e|m)sim(m,e)$, so typing information is only used implicitly via co-training; (2) multiply similarity with coarse, fine, or both typing scores. Note that the combination requires trivial additional computation for scoring. We experiment different typing score combinations in Table 4, evaluated on TAC-KBP2015. Combining only fine typing score, i.e. $P(e|m)sim(m,e)s_f(m,e)$, performs better among different settings.

All encoders in DE, TyDE and CA are initialized from the public Chinese BERT-base checkpoint. Details on model implementation and hyperparameters are in Appendix B.

5 Experiments

5.1 Evaluation on TAC-KBP2015

To compare our models with prior work, we benchmark on the established TAC-KBP2015 Chinese EL task [5]. Note that TAC-KBP2015 was originally designed for cross-lingual EL, but still suitable as a monolingual benchmark. Following De Cao et al. (2021b), we only evaluate in-KB links and do not consider NIL entities. We use full Chinese Wikipedia (E_{known} and E_{new}) as our target KB [6]. The evaluation metric is Recall@K, where R@1 is equivalent to accuracy (Botha et al., 2020).

[5]TAC-KBP2015 data is available at https://catalog.ldc.upenn.edu/LDC2019T02 and its license is at https://www.ldc.upenn.edu/data-management/using/licensing

Strategy	R@1
DE	75.2
TyDE (sim only)	75.9
TyDE (sim+coarse)	74.9
TyDE (sim+fine)	**76.2**
TyDE (sim+coarse+fine)	75.1

Table 4: Evaluations of TyDE inference strategy on TAC-KBP2015. We compare multiplying similarity with coarse, fine or both typing scores.

To be comparable with prior work, we use the published alias table from De Cao et al. (2021b) and the TAC-KBP2015 train set to extend *AT-base*, denoted as *AT-ext*. Models are trained with E_{known} examples only, as described in Section 3.2, where only *AT-base* was used for generating negatives. We further fine-tune CA on TAC-KBP2015's training set for 1 epoch, using *AT-ext* to generate negatives. The finetuned model is denoted as *CA-tuned*.

We evaluate DE, TyDE and CA models, based on *AT-ext*'s top-10 candidates. Table 3 shows evaluation results. Despite using a monolingual EL approach, our best model is on a par with the state-of-the-art model using multilingual data for training. In particular, CA-tuned outperforms all previous models with an XEL setting (Sil et al., 2018; Upadhyay et al., 2018). An error analysis for CA-tuned on TAC-KBP2015 is in Appendix F.

5.2 Evaluation on Hansel

We evaluate our models on Hansel-FS and Hansel-ZS, setting up a baseline for future work. When evaluating against Hansel, we do not use dataset-specific tuning. We use *AT-base* as the alias table and evaluate DE and CA based on *AT-base*'s top-10 candidates. Evaluation results of different systems on Hansel are shown in Table 5.

Comparison with mGENRE. To compare with prior work, we evaluate the state-of-the-art model mGENRE (with implementation details in Ap-

[6]We use a Freebase API to resolve predictions to a Freebase MID, to be consistent with the dataset. When our system cannot resolve the link, it counts as a prediction error.

	In-KB									With-NIL	
	AT			TyDE	CA	GEN.	+margin	+cand	+both	AT	CA+TyDE
Metric	R@1	R@10	R@100	R@1	R@1	R@1	R@1	R@1	R@1	R@1	R@1
Hansel-FS	0.0	61.1	63.0	11.7	**46.2**	36.6	35.2	35.2	35.6	0.0	44.1
Hansel-ZS	70.6	78.5	78.8	71.6	**76.6**	67.9*	66.8*	68.4*	68.4*	63.0	70.7

Table 5: Evaluation of our baselines and mGENRE models (denoted as GEN.) on the Hansel dataset. Both datasets are challenging for the state-of-the-art MEL model, while our CA model generalizes better to few-shot and zero-shot settings. mGENRE numbers on Hansel-ZS*: does not follow zero-shot training constraints, but still lower than CA results.

pendix H). Table 5 shows the results. According to our experiment, the base version of mGENRE outperforms ones with candidates and marginalization. This may be due to the low recall of AT on the FS slice, while the base model can recover some AT misses. Our CA model outperforms mGENRE by a large margin (+9.6) on this dataset.

We also evaluate mGENRE on the zero-shot slice. Note that mGENRE was trained on a Wikidata dump that overlaps with E_{new}, partially violating the zero-shot constraint, but the best variant still under-performs CA (-8.7). The ZS slice appears easier than FS, as all examples in FS are unsolvable by AT@1 but there is no such constraint in our zero-shot data collection process. Particularly, the adversarial mentions in ZS can link to head entities.

In short, our CA model is currently the best-performing for both zero-shot (76.6%) and few-shot (46.2%) slices, outperforming mGENRE by a large margin on both scenarios. This suggests that CA is less prone to popularity bias and generalizes better to tail and emerging entities. Large room of improvement remains on both datasets.

Error analysis. We perform an analysis on CA errors on Hansel-FS. 75% errors do not have the mention-entity pair as a top-10 alias table entry, suggesting major headroom of overcoming the restriction of alias tables. Among a sample of 40 other errors, for 30% cases CA predicts a general entity where the ground truth (GT) is a more specific instance. 28% errors are confusion with locations. 15% are confusion with temporal attributes. 10% are where CA predicts an irrelevant specific entity where GT is more general. Detailed error examples for each bucket is given in Appendix G.

NIL typing. We also set a baseline for entity linking with NIL classification for Hansel. In this baseline, we use CA model to rank *AT-base*'s top-10 candidates and use TyDE model's coarse classification head to compute NIL type. A NIL output is predicted if there is no candidate with output probability above a threshold of 0.1. We classify CA's NIL output with TyDE coarse typing result, and report the results in Table 5 as the baseline.

6 Conclusion

To address the popularity and language bias with Entity Linking datasets, we present a new benchmark consisting of two parts: the few-shot (FS) slice where the correct entities are not the most popular, and the zero-shot (ZS) slice where the entities are not observed in training. We name our dataset Hansel as both slices are in simplified Chinese (zh-hans), and make eval sets as well as the processed training set publicly available. Along with the dataset, we propose a method to collect human-calibrated few-shot and zero-shot EL datasets.

To compare with prior work, we build baseline models including a dual-encoder (DE) model, a novel typing-enhanced dual-encoder model (TyDE), and a cross-attention scoring model (CA). All models are supervised by hyperlinks in Chinese Wikipedia, and we make sure that new entities in the zero-shot slice are not visible during training.

On the TAC-KBP2015 Chinese EL task, our CA model (tuned on task-specific training set) gets R@1 of 88.1%, outperforming previous works with Cross-Lingual EL settings, achieving competitive results with mGENRE, the state-of-the-art Multilingual EL (MEL) model. Our CA model is the best-performing monolingual model on the established benchmark. Our TyDE model improves over a standard DE with minimal added complexity.

On Hansel, mGENRE only achieves a R@1 of 36.6% on the FS slice, much lower than its performance on TAC-KBP2015, suggesting difficulty of our dataset. Our CA model has so far the best R@1 of 46.2% on Hansel-FS, and R@1 of 76.6% on Hansel-ZS, outperforming mGENRE on both slices by a large margin. Future work on Chinese or multilingual EL may use our benchmark to test generalization over tail and emerging entities.

7 Limitations

There are a few limitations of our work worth noting. First, though the data collection method is applicable to any language, this time we release a Chinese-only dataset to fill the vacancy in this language, and leave other non-English zero-shot and few-shot EL datasets for future work. To construct such datasets for a new language, we discuss the necessary steps in Section 3.6 using our proposed dataset collection method.

Second, the proposed model that works best on Hansel requires cross-encoding mention context and entity description, which is computationally expensive as every retrieved mention-entity pair goes through inference. Our experiments show that dual-encoder based approach under-perform on Hansel, so it remains a challenge to perform well on our dataset with more efficient implementations.

Potential Risks. This work aims at alleviating the English bias for EL rooted from underexposure for non-English languages in EL datasets (Botha et al., 2020), particularly for zero-shot and few-shot settings. A potential risk that remains is underexposure of other (non-English and non-Chinese) languages for this problem, which we leave for future work. Nevertheless, the dataset collection methodology proposed in our work makes a step towards creating multilingual zero-shot and few-shot datasets for EL.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA. USENIX Association.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2306–2317, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021b. Multilingual autoregressive entity linking. *arXiv preprint arXiv:2103.12528*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020a. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

Thibault Févry, Nicholas FitzGerald, Livio Baldini Soares, and Tom Kwiatkowski. 2020b. Empirical evaluation of pretraining strategies for supervised entity linking. In *Automated Knowledge Base Construction*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Megan Leszczynski Daniel Y Fu, Mayee F Chen, and Christopher Ré. Tabi: Type-aware bi-encoders for open-domain entity retrieval.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria. PMLR.

Xianpei Han, Zhichun Wang, Jiangtao Zhang, Qinghua Wen, Wenqi Li, Buzhou Tang, Qi Wang, Zhifan Feng, Yang Zhang, Yajuan Lu, et al. 2020. Overview of the ccks 2019 knowledge graph evaluation track: Entity, relation, event and qa. *arXiv preprint arXiv:2003.03875*.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 385–396, New York, NY, USA. Association for Computing Machinery.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Heng Ji, Joel Nothman, H Trang Dang, and Sydney Informatics Hub. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. *Proceedings of TAC*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *TAC*.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, Cash Costello, and Sydney Informatics Hub. 2017. Overview of tac-kbp2017 13 languages entity discovery and linking. In *TAC*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, San Diego, CA.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *arXiv preprint arXiv:2001.03765*.

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

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Vera Provatorova, Samarth Bhargav, Svitlana Vakulenko, and Evangelos Kanoulas. 2021. Robustness evaluation of entity disambiguation using prior probes: the case of entity overshadowing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10501–10510, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.

Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, pages 457–466.

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, Online. Association for Computational Linguistics.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 428–438. Springer.

Weixin Zeng, Xiang Zhao, Jiuyang Tang, Zhen Tan, and Xuqian Huang. 2020. Cleek: A chinese long-text corpus for entity linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2026–2035.

Ce Zhang, Christopher Ré, Amir Sadeghian, Zifei Shan, Jaeho Shin, Feiran Wang, and Sen Wu. 2014. Feature engineering for knowledge base construction. *IEEE Data Eng Bull*.

Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. Towards zero-resource cross-lingual entity linking. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 243–252. Association for Computational Linguistics.

A Few-Shot Slice Collection Details

We detail the process using the alias table *AT-base* to generate a diverse known slice.

Corpora. The FS slice is constructed from three corpora: LCSTS (Hu et al., 2015) covers Weibo microblogging short text. The dataset is available at http://icrc.hitsz.edu.cn/Article/show/139.html, under CC BY-NC license. We sample examples from PART-I of LCSTS. SohuNews and TenSiteNews cover long news articles, from Sohu website and other news sites in Chinese respectively (Wang et al., 2008). They are available at http://www.sogou.com/labs/resource/list_news.php, namely SogouCA and SogouCS datasets. License for the dataset is at http://www.sogou.com/labs/resource/license_en.php.

Alias matching. We apply the alias table to perform exact matching on each unlabeled corpus. During alias matching, we favor long mentions over short ones if multiple mentions overlap. We apply a few Chinese-specific design decisions: (1) heuristically filter out single-character mentions to reduce noise; (2) do not use any tokenization mechanism, since space-tokenization is not available in Chinese, and any tokenizer may introduce system bias. (3) also compute $P(unlinked|m)$, i.e. the prior of a given phrase that do not have a hyperlink in Wikipedia. We removed the mentions that are over-commonly missing hyperlinks in Wikipedia, defined by $P(unlinked|m) > 0.98$. We found that this empirically gives a much cleaner candidate set thus saving annotation efforts.

Mention sampling. The alias matching produces a large candidate set over each corpus, which is unfeasible to label thoroughly. To sample a diverse and representative subset, we take diverse mentions and documents into the sample. We sample each corpus by two equal criteria to get sets of mention phrases, then randomly select one example per phrase. The criteria are namely (1) uniformly sample, and (2) sample only ambiguous mentions with at least two candidates in the alias table.

Handling offensive or sensitive data. During annotations both for FS and ZS slices, we asked annotators to remove an example if it contains offensive information or sensitive data that might uniquely identify individual people.

As shown in Table 1, Hansel-FS features a diverse set of 2.7K entities from 5.2K different documents.

B Experiment Details

We implement DE, TyDE and CA models using Tensorflow (Abadi et al., 2016). The DE, TyDE and CA encoders all use 12 transformer encoder layers, initialized with BERT-base parameters. The number of parameters for DE, TyDE and CA are roughly 204M, 210M, 102M. We use Adam optimizer (Kingma and Ba, 2015) with linear weight decay and use 10% steps for a linear warmup schedule, following Botha et al. (2020).

The models are trained on a single NVIDIA V100 GPU. All general models are trained for 100K steps. Training of DE and TyDE model takes approximately 30 hours. Training CA on Wikipedia takes 16 hours, and finetuning CA on TAC-KBP2015 takes 4 hours. Every reported result is from a single run.

We fix sequence length to be 128 tokens for both mention and entity encoder for DE and TyDE, and 256 tokens for CA. We select the approximate maximum batch size that fits into the GPU memory, resulting in a batch size of 64 for DE and TyDE, and 32 for CA. We search learning rate among [1e-5, 2e-5, 1e-4] for DE and TyDE. Following Botha et al. (2020), we fix 1e-5 as the learning rate for CA. We search learning rate among [1e-6, 5e-6] for CA-tuned. We search mention and entity embedding dimension d within [128, 256] for DE and TyDE. We perform one hyper-parameter search, using batch accuracy in validation set for DE and TyDE and classification accuracy for CA to make hyper-parameter choices. Best-performing hyper-parameters are: learning rate is 2e-5 for DE and TyDE, and 5e-6 for CA-tuned. Embedding dimension d is 256. We choose 0.1 as the NIL threshold probability for CA+TyDE model, for With-NIL evaluations.

C TopSnaks Examples

Table 6 shows 40 examples of Wikidata TopSnaks from the 2018 dump. From the table we see that TopSnaks include diverse entity attributes such as types, gender, occupation, country and sport. Intuitively, our TyDE models encourage the learned mention and entity embeddings to capture rich information supervised by these TopSnaks.

D Wikidata Filtering

Following a similar constraint with Botha et al. (2020), when processing Wikidata dumps, we fil-

TopSnak	Snak name
P31-Q13442814	instance of: scholarly article
P31-Q5	instance of: human
P21-Q6581097	sex or gender: male
P31-Q16521	instance of: taxon
P105-Q7432	taxon rank: species
P17-Q148	country: People's Republic of China
P421-Q6985	located in time zone: UTC+08:00
P17-Q30	country: United States of America
P31-Q7187	instance of: gene
P21-Q6581072	sex or gender: female
P17-Q145	country: United Kingdom
P407-Q1860	language of work or name: English
P31-Q13100073	instance of: village-level division in China
P279-Q20747295	subclass of: protein: coding gene
P31-Q8054	instance of: protein
P17-Q183	country: Germany
P31-Q8502	instance of: mountain
P279-Q8054	subclass of: protein
P31-Q486972	instance of: human settlement
P106-Q82955	occupation: politician
P279-Q7187	subclass of: gene
P17-Q142	country: France
P31-Q4022	instance of: river
P641-Q2736	sport: association football
P17-Q159	country: Russia
P27-Q30	country or citizenship: USA
P1435-Q15700834	heritage designation: Grade II listed building
P17-Q55	country: Netherlands
P31-Q79007	instance of: street
P17-Q20	country: Norway
P31-Q3305213	instance of: painting
P31-Q54050	instance of: hill
P17-Q16	country: Canada
P421-Q6723	located in time zone: UTC+02:00
P31-Q532	instance of: village
P17-Q34	country: Sweden
P31-Q17329259	instance of: encyclopedic article
P407-Q7737	language of work or name: Russian
P17-Q96	country: Mexico
P421-Q6655	located in time zone: UTC+01:00

Table 6: Example TopSnaks.

Types	QIDs
Disambiguation page	Q4167410
Templates	Q11266439 Q105528595 Q11753321 Q15671253 Q19887878 Q20769160 Q24731821 Q26142649 Q26267864 Q36330215 Q4657797 Q48552277 Q56876519 Q74980542 Q95691391 Q97303168
Categories	Q4167836 Q105653689 Q13406463 Q1474116 Q15407973 Q15647814 Q20769287 Q24574745 Q30432511 Q54662266 Q59542487 Q56428020
Modules	Q15184295 Q15145755 Q18711811 Q59259626
Wikimedia project page	Q14204246
Subclasses of above	Q97011660
	Q11266439 Q25051296 Q21528878 Q4663903 Q13406463 Q22247630 Q30415057 Q60715851 Q15184295

Table 7: WikiData identifiers used for filtering out Wikimedia-internal entities.

tered out entities that are a subclass (P279) or instance of (P31) Wikimedia-internal administrative entities. We extended the list of such entities by Botha et al. (2020), detailed in Table 7.

E More Examples of Hansel

In Table 8, we provide examples of Hansel-FS Slice along with CA model predictions, to demonstrate properties of the dataset and model. From the analysis, we see that the CA model can capture information in types and relations (e.g. "Line 13" and "Qu Bo" examples), while also making some mistakes with entities with similar types or meaning (see the tennis example). It also demonstrates that Hansel-FS is a challenging benchmark.

In Table 9, we provide examples of Hansel-ZS to demonstrate its properties. As shown in the examples, our annotation process cultivated some genuinely hard and ambiguous zero-shot examples.

F Error Analysis for CA-tuned on KBP-2015

We do a brief error analysis on CA-tuned results on TAC-KBP2015. Among all R@1 errors, 212 (19%) do not have a Chinese Wikipedia page. Note that we constrain our model to a monolingual setting thus missing these examples, whereas Cross-Lingual and Multilingual models (Upadhyay et al., 2018; De Cao et al., 2021b) are inherently better at such examples. 544 (48%) errors do not have the mention-entity pair in top-10 alias table entries, indicating headroom of retrieval or generation models without reliance on alias tables. 344 (30%) cases are where the model did not choose the correct candidate. In 39 (3.4%) cases the freebase MIDs are not resolved to Wikidata.

G Error Analysis of CA on Hansel-FS

We provide detailed examples in Table 10 and Table 11 for CA model's common prediction errors on the challenging Hansel-FS dataset. Specifically, we did not include alias table misses, and for the rest of the errors, we categorize 40 examples into buckets and visualize the top-4 major buckets. Overall, predicting a common entity while a specific one exists, or predicting a specific entity while a common one is more suitable, are two common error categories. This suggests that a better context comprehension for entities' granularity may be needed. There are also common errors confusing entities with different geographical and temporal attributes, indicating that a better utilization of entity attributes and fine-grained types may be required, in order to improve results on the dataset.

H mGENRE Implementation Details

We use the code in the publicly available GENRE repository[7]. We use the provided mGENRE model and do not perform any fine-tuning to its parameters. Since mGENRE uses both Wikipedia and Wikidata dumps from 2019-10-01, and our ZS slice include entities from Wikidata 2021-03-15, for Hansel-ZS evaluations, we extend the catalog of entity names by considering all languages for each entity from E_{new}, obtained from the Wikidata dump.

I Comparision of Existing Chinese EL Datasets and Hansel

The only 2 series of Chinese EL datasets that link to Wikidata are TAC-KBP series (Ji et al., 2015, 2016, 2017) and CLEEK (Zeng et al., 2020). Table 12 summarizes the datasets' statistics and domains. Our dataset sets itself apart by filling the vacancy of non-English few-shot and zero-shot challenges.

To obtain a few-shot slice, it is intuitive to subsample TAC-KBP or CLEEK, i.e. removing correct AT@1 as we do in the human annotation stage. Although sub-sampling is feasible, its major disadvantage is the lack of mention and entity diversity. As Table 12 shows, the subsets of TAC-KBP and CLEEK, after removing correct AT@1 examples, lack diversity due to their intrinsic features. Take TAC-KBP2017 for example, its few-shot subset has 3,883 mentions, covering only 877 different surface forms, 167 documents and 350 entities, suggesting lots of lexical repetitions across examples. On the other hand, Hansel-FS has 5,260 (1.4x) mentions, covering 4,097 (5x) different surface forms, 5,234 (30x) documents and 2,720 (8x) entities. The diversity of Hansel-FS is rooted from our collection method, as we sample mentions from a large set of documents, avoiding repetitive mentions and entities that commonly appear in a same document, making the dataset challenging and syntactically diverse.

In the zero-shot slice, we use the emerging entities in temporally evolving Wikidata to collect Hansel-ZS. We apply this zero-shot setting due to

[7] https://github.com/facebookresearch/GENRE

Context	… 2013赛季意大利BNL国际赛继续进行，3号种子阿扎伦卡在6-1/2-0领先时收到了森田步美的退赛大礼，顺利晋级八强的白俄罗斯人将在 [E1] 1/4决赛 [/E1] 迎战9号种子斯托瑟，后者在另一场比赛中以7-5/2-6/6-1首次击败科维托娃，结束了对捷克人的四连败 …
Translation	The 2013 Italian Open continued. No. 3 seed Azarenka received the message of Ayumi Morita's withdrawal when taking a 6-1/2-0 lead. The Belarusian athlete successfully advanced to the **[E1] quarter finals [/E1]** …
Prediction	`2010`年澳洲网球公开赛男子单打比赛[Q782589]: 拉斐尔·纳达尔是上届冠军，但在半决赛败给当届亚军安迪·穆雷。 最终冠军为罗杰·费德勒，决赛以6-4、6-3、7-6直落三盘击败安迪·穆雷 …
Translation	`2010 Australian Open - Men's Singles`[Q782589]: Rafael Nadal was the last champion, but lost to current runner-up Andy Murray in the semifinals. The final champion was Roger Federer. In the final …
Annotation	NIL_EVENT
Outcome	**Wrong**: "Quarter-finals" is a rare alias of predicted entity "2010 Australian Open–Men's Singles" (a tennis tournament). Although the context is relevant to tennis, it should refer to a different tournament in 2013 not in the KB. The model sniffs for an irrelevant entity with a similar type.
Context	据北京地铁官方微博5日早晨7时55分发布消息称，目前， [E1] 地铁13号线 [/E1] 因设备故障，影响部分列车晚点，运行间隔较大，建议有急事的乘客请选择其他交通工具 …
Translation	According to the Beijing Metro official Weibo at 7:55 on the 5th, at present, **[E1] Metro Line 13 [/E1]** has some equipment failures, and some trains are delayed and the operation interval is relatively long. Passengers in urgent matters are advised to please Choose other transportation …
Prediction	北京地铁`13`号线[Q847744]: 北京地铁13号线，又称北京城市铁路，简称"城铁"，旧称北京轻轨铁路，又常被误称为"轻轨"，是一条连接中国北京市西城区西直门站至东城区东直门站，属于北京地铁的地铁线路…
Translation	`Beijing Metro Line 13`[Q847744]: Beijing Metro Line 13, also known as Beijing Urban Railway, or "Urban Rail" for short. It was formerly known as Beijing Light Rail Railway and often mistakenly called "Light Rail". It is a line connecting Xizhimen Station in Xicheng District, Beijing, China to Dongzhimen …
Outcome	**Correct**: Although the mention "line 13" is ambiguous, the model correctly resolves the mention to that in Beijing (not the top candidate in the alias table), as is sniffs 'Xizhimen Station" in the context, a station on the Beijing subway line 13.
Context	7月1日晚7点半，中超联赛第15轮贵州茅台队VS大连实德队的比赛在贵阳奥体中心点燃战火。凭借 [E1] 曲波 [/E1] 第5分钟和穆斯利第76分钟的进球，贵州茅台队最终2：0击败大连实德，本赛季首获3连胜 …
Translation	At 7:30 pm on July 1, the 15th round of the Chinese Super League match between Kweichow Moutai vs. Dalian Shide ignited at the Guiyang Olympic Sports Center. With **[E1] Qu Bo [/E1]** scored in the 5th minute and Mooseley scored in the 76th minute, Kweichow Moutai defeated Dalian Shide 2-0 and won the first three consecutive victories this season. …
Prediction	曲波[Q483636]：曲波，出生于天津，已退役的中国足球运动员，曾是中国国家足球队成员 …
Translation	`Qu Bo`[Q483636]: Qu Bo, born in Tianjin, a retired Chinese football player who was a member of the Chinese national football team …
Outcome	**Correct**: The context is about soccer, and the model correctly links the name "Qu Bo" to the soccer player even though it is not the most popular candidate in the alias table.

Table 8: Examples in Hansel-FS slice with CA model predictions.

its practical use. Since EL is often used in knowledge base construction and population (Shen et al., 2014; Hoffart et al., 2014), this setting simulates how to link mentions to emerging entities with 2018's training data.

The TAC-KBP datasets are available for a price. For the convenience of future research, Hansel is open-source, including the processed Wikipedia dump as training data, to better facilitate the development of new Chinese EL systems.

In conclusion, Hansel-FS and Hansel-ZS provide a robust and comprehensive benchmark on Chinese EL systems and cannot be substituted by simply subsampling existing datasets.

J Ethical Considerations

In this section, we discuss the main ethical considerations of Hansel:

Intellectual property protection. Hansel's training data is processed from Wikipedia and most of Wikipedia' text is licensed under CC BY-SA. The original text of Hansel-FS is from LCSTS (Hu et al., 2015), SohuNews and TenSiteNews (Wang et al., 2008). LCSTS grants the permission to copy,

Mention 1	来源：新闻晨报 记者：王嫣 今天上午，2019年 [E1] 上海大师赛 [/E1] 举行了男单正赛的抽签仪式。两届大满贯冠军、今年进入网球名人堂的李娜与获得男单正赛外卡的张之臻 . . .
Translation	Source: Morning Post. Reporter: Yan Wang. This morning, the draw ceremony of the men's singles competition was held in the 2019 **[E1] Shanghai Masters [/E1]**. Na Li, who won the Grand Slam champion twice and entered the Tennis Hall of Fame this year, together with Zhizhen Zhang, who won . . .
Entity 1	`2019年上海大师赛`[Q69355546]: 2019年上海大师赛为第12届上海大师赛，又名2019年上海劳力士大师赛，是ATP世界巡回赛1000大师赛事的其中一站 . . .
Translation	`2019 Shanghai Masters`[Q69355546]: The 2019 Shanghai Masters, also known as the 2019 Shanghai Rolex Masters, was the 12th Edition of the Shanghai Masters, classified as an ATP Tour Masters . . .
Mention 2	#2020斯诺克世锦赛# 交手记录 . . . 2017年英格兰公开赛决赛：奥沙利文9-2威尔逊 2018年 [E1] 上海大师赛 [/E1] 半决赛：奥沙利文10-6威尔逊 2018年"冠中冠"邀请赛决赛：奥沙利文10-9威尔逊 . . .
Translation	#2020 World Snooker Championship# Match Record . . . 2017 English Open Final: O'Sullivan 9-2 Wilson 2018 **[E1] Shanghai Masters [/E1]** Semi-final: O'Sullivan 10-6 Wilson 2018 Champion of Champions . . .
Entity 2	`2019年斯诺克上海大师赛`[Q66436641]: 2019年世界斯诺克·上海大师赛属职业斯诺克非排名赛，于2019年9月9日－15日在上海富豪环球东亚酒店举行。 . . .
Translation	`2019 Shanghai Snooker Masters`[Q66436641]: The 2019 World Snooker Shanghai Masters was a professional non-ranking snooker tournament that took place at the Regal International East Asia Hotel . . .
Mention 3	这是2019年11月30日 [E1] 上海大师赛 [/E1] "传奇赛"对决的决赛，中国的传奇队是来自退役选手Gogoing、Melon、小伞、U和诺夏组成OMG的班底，而他们的对手则是韩国的退役选手。 . . .
Translation	This is the final of "Legend Tournament" on **[E1] Shanghai Masters [/E1]** on November 30, 2019. The legendary team of China is a team of retired players, consisting of Gogoing, Melon, Xiaosan, U and Nuoxia from OMG Organization. Their opponents are retired players from South Korea . . .
Entity 3	NIL_EVENT
Analysis	During data collection, Entity 1 (entity in E_{new}) was provided. The annotator found Mention 1 via Web search, as well as two adversarial mentions with the same phrase ("Shanghai Masters"), referring to a tennis tournament, a snooker tournament, and an online gaming tournament respectively.
Mention 1	1905电影网讯 已经筹备了十余年的吉尔莫·德尔·托罗的《 [E1] 匹诺曹 [/E1] 》，在上个月顺利被网飞公司买下，成为了流媒体巨头旗下的新片。 . . .
Translation	(1905 Film Network News) Having prepared for more than 10 years, Guillermo del Toro's **[E1] Pinocchio [/E1]** was successfully acquired by Netflix, becoming a new film of the streaming media giant . . .
Entity 1	`木偶奇遇记_(2021年电影)`[Q73895818]: 《木偶奇遇记》（暂名，）是一部预定于2021年上映的美国3D定格动画黑暗奇幻歌舞片，由吉勒摩·戴托罗执导。 . . .
Translation	*`The Adventures of Pinocchio`*`_(2021 film)`[Q73895818]: *The Adventures of Pinocchio* (tentative name) is an upcoming American stop-motion animated dark fantasy musical film directed by Guillermo del Toro and is planned for a 2021 release . . .
Mention 2	[E1] 匹诺曹 [/E1] 的金币还是被狐狸和猫骗走了。他去报官，发现猴子法官说话颠三倒四，喜欢抓无辜的人。无奈之下，匹诺曹只好编造谎言，说自己偷了很多东西了，最终才得以逃离。 . . .
Translation	The fox and the cat swindled **[E1] Pinocchio [/E1]** out of his coins. Pinocchio went to report to the officials and found that the Monkey Judge talked incoherently and liked to catch innocent people. In desperation, Pinocchio had no choice but to fabricate a lie, claiming that he had stolen tons of things, and finally escaped.
Entity 2	匹诺曹[Q6502703]: 匹诺曹，名字来自意大利语""（"松果"），是一个虚构人物，意大利作家卡洛·科洛迪所着儿童文学作品《木偶奇遇记》（1883年）的主角，在原版同时也是反派角色之一 . . .
Translation	`Pinocchio`[Q6502703]: Pinocchio, whose name comes from the Italian words *pino* (pine), is a fictional character and the protagonist of the children's novel The Adventures of Pinocchio (1883) by Italian writer Carlo . . .
Mention 3	#匹诺曹定档#改编自经典童话《木偶奇遇记》的奇幻电影《 [E1] 匹诺曹 [/E1] 》发布定档预告，定档6月1日儿童节。影片由马提欧·加洛尼（《犬舍惊魂》）执导，罗伯托·贝尼尼（《美丽人生》 . . .
Translation	#*Pinocchio*ReleaseDate# The fantasy film "**[E1] Pinocchio [/E1]**", adapted from the classic fairy tale, will be released on June 1st for Children's Day. The film is directed by Matteo Galloni ("The Kennel") . . .
Entity 3	NIL_OTHER
Analysis	All with the same mention text, Mention 1 refers an entity in E_{new} which is a 2021 film directed by G. del Toro, with a different canonical name than the mention. Mention 3 refers to another film Pinocchio in 2019 by M. Garrone, which is not in zh-wiki thus deserves a NIL label. Mention 2 refers to the fictional character.

Table 9: Examples in Hansel-ZS slice, illustrating challenging zero-shot and adversarial examples collected by annotators.

Bucket 1	**Predicted general entity while specific one exists (30%)**
Context	… 英国威廉王子办公室宣布，威廉王子的妻子凯特王妃已怀有身孕。办公室在声明中称：“[E1] 剑桥公爵 [/E1] 及公爵夫人非常高兴地宣布，公爵夫人已有喜。”网友们也不甘寂寞，合成了未来宝宝的样子，宝宝相貌神似父母，趣味十足。 …
Translation	…The office of Prince William of England announced that Prince William's wife, Princess Kate, is pregnant. The office said in a statement: “**[E1] The Duke of Cambridge [/E1]** and the Duchess are very happy to announce that the Duchess has been happy.” Netizens were unwilling to be lonely, and synthesized the appearance of the future baby. The baby looks like his parents and is full of fun. …
Prediction	剑桥公爵[Q836810]：剑桥公爵（又译坎布里奇公爵）为其中一种，也是英国王室的一种特别等级。此头衔（以英格兰剑桥为名）可经由长子继承制，由男性后裔继承，并已授予多位英国王室成员。剑桥公爵的配偶则称作剑桥公爵夫人。 …
Translation	`The Duke of Cambridge`[Q836810]: The Duke of Cambridge (also translated as the Duke of Cambridge) is one of them, and it is also a special rank of the British royal family. This title (under the name of Cambridge, England) can be inherited by male descendants through the eldest son inheritance system …
Annotation	剑桥公爵威廉王子[Q36812]：剑桥公爵威廉王子殿下，全名为威廉·亚瑟·菲利普·路易，是王储威尔斯亲王查尔斯与威尔斯王妃戴安娜的长子，英国女王伊丽莎白二世与菲利普亲王的长孙。 …
Translation	`Prince William, Duke of Cambridge`[Q36812]: His Royal Highness Prince William, Duke of Cambridge, whose full name is William Arthur Philip Louis, is the eldest son of Prince Charles of Wales and Diana, Princess of Wales, and the eldest grandson of Queen Elizabeth II and Prince Philip of England. …

Bucket 2	**Predicted similar entity with wrong location (28%)**
Context	… “当时我站在大盆旁边，等着衣服被甩干，没想到衣服刚刚放进没有一分钟，洗衣机爆炸了。碎片一院子飞的都是，连厨房里也蹦进了不少碎片，还好儿子没事，不过现在想想还是后怕。”家住 [E1] 市中区 [/E1] 西王庄乡民主村的村民邵艳伟说。 …
Translation	…“I was standing next to the big basin, waiting for the clothes to be dried. I didn't expect that the washing machine exploded within a minute after the clothes were put in. The debris was flying all over the yard, and even a lot of debris jumped into the kitchen. My good son is okay, but I'm still scared when I think about it now.” said Shao Yanwei, a villager who lives in **[E1] Shizhong District [/E1]** Xiwangzhuang Township Democracy Village. …
Prediction	市中区[Q598098]：市中区是中国山东省济南市所辖的市辖区，这个区面积为280平方公里，人口总数为57万人（2004年）。 …
Translation	`Shizhong District`[Q598098]: Shizhong District is a municipal district under the jurisdiction of Jinan City, Shandong Province, China. This district covers an area of 280 square kilometers and has a total population of 570,000 (2004). …
Annotation	市中区[Q1198415]：市中区是中国山东省枣庄市所辖的一个市辖区。总面积为375平方千米，2001年人口为48万。 …
Translation	`Shizhong District`[Q1198415]: Shizhong District is a municipal district under the jurisdiction of Zaozhuang City, Shandong Province, China. The total area is 375 square kilometers, and the population in 2001 was 480,000. …

Table 10: Error analysis of CA model on Hansel-FS slice. (Bucket 1 and 2)

distribute and modify under the terms of CC BY-NC License. The SohuNews and TenSiteNews's license grants the permission to carry out research or study to form achievement with its own intellectual property rights. Hansel-ZS is collected with searching-based annotation. Hence all data in this slice is in public domain.

Annotation participants and payments. Participants are 15 undergraduate students with Chinese as their native language, who major in computer science and have basic understanding of entity linking. They are all well aware of how the collected data will be used. The salary for annotating each entry is determined by the average time of annotation and local labor compensation standard.

Annotation Instructions. The instructions are explicitly given in the annotation interface. Figure 4 is a screenshot of Hansel-FS annotation. Figure 5 and Figure 6 are screenshots of Hansel-ZS annotation.

Bucket 3	**Similar entity with wrong date (15%)**
Context	… ４月２９日，王一梅右脚脚踝韧带撕裂，并经历了手术治疗；７月１日，伤愈归队；７月２０日，主帅俞觉敏曾向记者介绍，大梅已恢复了五成功力……现在，王一梅已经随中国女排来到伦敦奥运会赛场。…“不过，毕竟手术到现在只有３个月，特别是王一梅归队之后与队伍的整体磨合只有１０天，时间非常紧，到了 [E1] 奥运会 [/E1] 赛场上，她到底能发挥出怎样的状态，现在大家都没底……”至于昨天同英国女排的热身赛，俞觉敏直言，这同奥运会的正式比赛有着明显的不同 …
Translation	…On April 29, Wang Yimei suffered a torn ligament in her right ankle and underwent surgical treatment; on July 1, he returned to the team from injury; on July 20, coach Yu Juemin introduced to reporters that Damei had recovered his five strengths... Now, Wang Yimei has accompanied the Chinese women's volleyball team to the London Olympics. …The time is very tight. In the **[E1] Olympic Games [/E1]**, how can she perform? Nobody has any idea.” As for the warm-up match with the British women's volleyball team yesterday, Yu Juemin bluntly said that this is obviously different from the official Olympic game. …
Prediction	第二十九届现代夏季奥林匹克运动会[Q8567]：第二十九届现代夏季奥林匹克运动会，又称2008年夏季奥运会或北京奥运会，于2008年8月8日至24日在中华人民共和国首都北京举行。 …
Translation	`The 29th Modern Summer Olympic Games`[Q8567]: The 29th Modern Summer Olympic Games, also known as the 2008 Summer Olympics or Beijing Olympics, was held from August 8 to 24, 2008 in Beijing, the capital of the People's Republic of China. …
Annotation	2012年夏季奥林匹克运动会[Q8577]：2012年夏季奥林匹克运动会，正式名称为第三十届夏季奥林匹克运动会，又称为2012年伦敦奥运会，是于2012年7月27日至8月12日在英国伦敦举行的一届综合性运动会。 …
Translation	`The 2012 Summer Olympic Games`[Q8577]: The 2012 Summer Olympic Games, officially known as the 30th Summer Olympic Games, also known as the 2012 London Olympics, is a comprehensive sports meeting held in London, England from July 27 to August 12, 2012. …

Bucket 4	**Predicted an irrelevant specific instance of a general entity (10%)**
Context	… 中新网６月２８日电　据俄新网２７日报道，俄罗斯总统普京表示，通过直接投票的方式选举产生俄联邦委员会参议员的做法违反宪法，但是他不排除将来可能 [E1] 修改宪法 [/E1] 直接选举产生参议员。普京强调，“宪法规定，联邦委员会由执行和立法机关代表组成。”他指出，现行宪法没有规定选民直接投票选举产生参议员的程序。再被问及是否会为实现直接选举联邦委员会成员而修改宪法时，普京表示，“我不认为在这种情况下我们应该现在着手这个问题。但这在将来是有可能的。” …
Translation	…Chinanews.com, June 28. According to a report on the 27th of Russia's new website, Russian President Vladimir Putin stated that the election of senators to the Russian Federation Council through direct voting violates the Constitution, but he does not rule out the possibility of **[E1] amending the constitution [/E1]** in the future. Directly elected senators. Putin emphasized, “The Constitution stipulates that the Federal Council is composed of representatives of the executive and legislative bodies.” He pointed out that the current Constitution does not provide for the procedure for voters to directly vote for the election of senators. When asked again whether he would amend the constitution to achieve direct election of members of the Federal Council, Putin said, “I don't think we should tackle this issue now under such circumstances. But it is possible in the future.” …
Prediction	2020年俄罗斯修宪公投[Q598098]：2020年俄罗斯修宪公投是俄罗斯于2020年6月25日至7月1日举行的公投。此次公投是俄罗斯总统普京在2020年1月15日向联邦会议时提出的 …
Translation	`The 2020 Russian constitutional amendment referendum`[Q83347039]: The 2020 Russian constitutional amendment referendum is a referendum held by Russia from June 25 to July 1, 2020. The referendum was proposed by Russian President Vladimir Putin at the Federal Conference on January 15, 2020. …
Annotation	宪法修正[Q1198415]：宪法修正，简称修宪，指的是国家宪法的修改。有一些国家允许修改宪法本文；也有一些国家不能修改宪法本文，但允许在本文后面附上增修条文。 …
Translation	`Constitutional amendment`[Q53463]: Constitutional amendment, referred to as constitutional amendment, refers to the amendment of the national constitution. Some countries allow amendments to the text of the constitution; some countries cannot amend the text of the constitution, but allow additions and amendments to the back of the text. …

Table 11: Error analysis of CA model on Hansel-FS slice. (Bucket 3 and 4)

Dataset	#Mentions			#Distinct Mentions			#Documents			#Entities	Domains
	In-KB	NIL	Total	In-KB	NIL	Total	In-KB	NIL	Total		
TAC-KBP2015 (Ji et al., 2015)	8,666	2,400	11,066	1,246	1,627	2,869	166	146	166	840	News, Discussion Forum
TAC-KBP2016 (Ji et al., 2016)	7,115	1,730	8,845	1,185	1,080	2,221	166	167	167	742	News, Discussion Forum
TAC-KBP2017 (Ji et al., 2017)	7,673	2,573	10,246	1,218	1,297	2,421	167	167	167	796	News, Discussion Forum
CLEEK (Zeng et al., 2020)	2,609	177	2,786	1,435	135	1,569	100	55	100	1,191	News
TAC-KBP2015 FS Subset	2,072	316	2,388	417	140	555	155	90	161	298	News, Discussion Forum
TAC-KBP2016 FS Subset	2,255	581	2,836	475	241	679	166	130	167	354	News, Discussion Forum
TAC-KBP2017 FS Subset	2,583	1,300	3,883	486	464	877	163	159	167	350	News, Discussion Forum
CLEEK FS Subset	685	47	732	421	36	456	94	24	95	377	News
Hansel-FS (ours)	3,404	1,856	5,260	2,654	1,606	4,097	3,389	1,850	5,234	2,720	News, Social Media
Hansel-ZS (ours)	4,208	507	4,715	3,981	468	4,222	4,200	507	4,704	4,046	News, Social Media, E-books, etc.

Table 12: **Comparision of existing Chinese EL datasets and the Hansel dataset.** We break down the number of mentions, distinct mentions and documents by whether the label is a NIL entity or inside Wikidata (In-KB). We also provide statistics of existing datasets' few-shot (FS) subsets.

Figure 4: **Screenshot for Hansel-FS annotation.** Annotators are given a highlighted mention and its context and some possible choices to facilitate annotation. Detailed annotation procedure can be found in Section 3.3.

Figure 5: **Screenshot for Hansel-ZS annotation (Stage 1).** Annotators are given a entity and its basic information (i.e. entity name, aliases and description). Links for searching Weibo, Baidu News, etc. are provided to facilitate annotation.

Figure 6: **Screenshot for Hansel-ZS annotation (Stage 2).** Based on stage 1, annotators seek adversarial examples by searching for a same or similar mention referring to a different entity. Annotators may choose to repeat this stage to add multiple adversarial examples.