Zero-Shot Entity Linking by Reading Entity Descriptions

Lajanugen Logeswaran[†*] **Ming-Wei Chang**[‡] **Kenton Lee**[‡] **Kristina Toutanova**[‡]
Jacob Devlin[‡] **Honglak Lee**[‡,†]

[†]University of Michigan, [‡]Google Research

{llajan,honglak}@umich.edu,

{mingweichang,kentonl,kristout,jacobdevlin,honglak}@google.com

Abstract

We present the *zero-shot entity linking* task, where mentions must be linked to unseen entities without in-domain labeled data. The goal is to enable robust transfer to highly specialized domains, and so no metadata or alias tables are assumed. In this setting, entities are only identified by text descriptions, and models must rely strictly on language understanding to resolve the new entities. First, we show that strong reading comprehension models pre-trained on large unlabeled data can be used to generalize to unseen entities. Second, we propose a simple and effective adaptive pre-training strategy, which we term *domain-adaptive pre-training* (DAP), to address the domain shift problem associated with linking unseen entities in a new domain. We present experiments on a new dataset that we construct for this task and show that DAP improves over strong pre-training baselines, including BERT. The data and code are available at https://github.com/lajanugen/zeshel.[1]

1 Introduction

Entity linking systems have achieved high performance in settings where a large set of disambiguated mentions of entities in a target entity dictionary is available for training. Such systems typically use powerful resources such as a high-coverage alias table, structured data, and linking frequency statistics. For example, Milne and Witten (2008) show that by only using the prior probability gathered from hyperlink statistics on Wikipedia training articles, one can achieve 90% accuracy on the task of predicting links in Wikipedia test articles.

While most prior works focus on linking to general entity databases, it is often desirable to link to specialized entity dictionaries such as legal cases, company project descriptions, the set of characters in a novel, or a terminology glossary. Unfortunately, labeled data are not readily available and are often expensive to obtain for these specialized entity dictionaries. Therefore, we need to develop entity linking systems that can generalize to unseen specialized entities. Without frequency statistics and meta-data, the task becomes substantially more challenging. Some prior works have pointed out the importance of building entity linking systems that can generalize to unseen entity sets (Sil et al., 2012; Wang et al., 2015), but adopt an additional set of assumptions.

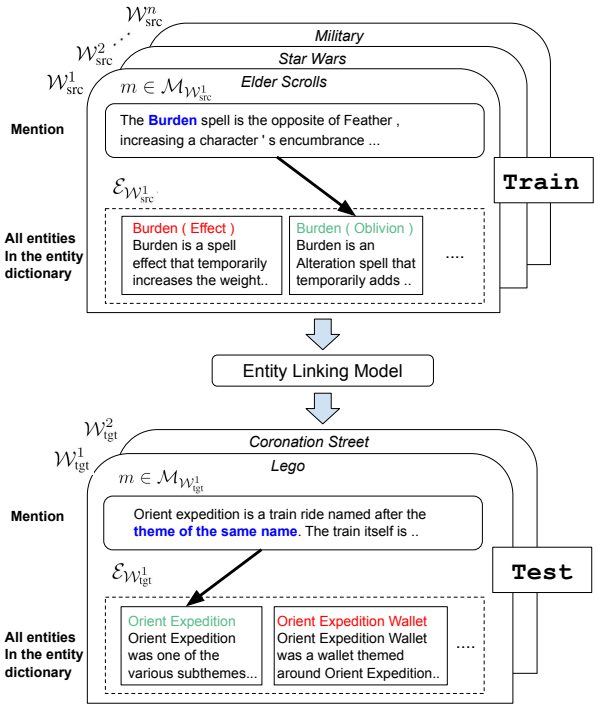

Figure 1: **Zero-shot entity linking.** Multiple training and test domains (worlds) are shown. The task has two key properties: (1) It is **zero-shot**, as no mentions have been observed for any of the test world entities during training. (2) Only **textual** (non-structured) information is available.

[*] Work completed while interning at Google

[1] zeshel stands for **ze**ro-**sh**ot **e**ntity **l**inking.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3449–3460
Florence, Italy, July 28 - August 2, 2019.

In this work, we propose a new *zero-shot entity linking* task, and construct a new dataset for it.[2] The target dictionary is simply defined as a set of entities, each with a text description (from a canonical entity page, for example). We do not constrain mentions to named entities, unlike some prior work, which makes the task harder due to large number of candidate entities. In our dataset, multiple entity dictionaries are available for training, with task performance measured on a disjoint set of test entity dictionaries for which no labeled data is available. Figure 1 illustrates the task setup. We construct the dataset using multiple sub-domains in Wikia and automatically extract labeled mentions using hyper-links.

Zero-shot entity linking poses two challenges for entity linking models. First, without the availability of powerful alias tables or frequency priors, models must read entity descriptions and reason about the correspondence with the mention in context. We show that a strong reading comprehension model is crucial. Second, since labeled mentions for test entities are not available, models must adapt to new mention contexts and entity descriptions. We focus on both of these challenges.

The contributions of this paper are as follows:

- We propose a new *zero-shot entity linking* task that aims to challenge the generalization ability of entity linking systems with minimal assumptions. We construct a dataset for this task, which will be made publicly available.
- We build a strong baseline by using state-of-the-art reading comprehension models. We show that attention between mention in context and entity descriptions, which has not been used in prior entity linking work, is critical for this task.
- We propose a simple yet novel adaptation strategy called domain-adaptive pre-training (DAP) and show that it can further improve entity linking performance.

2 Zero-shot Entity Linking

We first review standard entity linking task definitions and discuss assumptions made by prior systems. We then define the zero-shot entity linking task and discuss its relationship to prior work.

[2]Existing datasets are either unsuitable or would have to be artificially partitioned to construct a dataset for this task.

2.1 Review: Entity linking

Entity linking (EL) is the task of grounding entity mentions by linking them to entries in a given database or dictionary of entities. Formally, given a mention m and its context, an entity linking system links m to the corresponding entity in an **entity set** $\mathcal{E} = \{e_i\}_{i=1,\ldots,K}$, where K is the number of entities. The standard definition of EL (Bunescu and Pasca, 2006; Roth et al., 2014; Sil et al., 2018) assumes that mention boundaries are provided by users or a mention detection system. The entity set $\mathcal{E}$ can contain tens of thousands or even millions of entities, making this a challenging task. In practice, many entity linking systems rely on the following resources or assumptions:

Single entity set This assumes that there is a single comprehensive set of entities $\mathcal{E}$ shared between training and test examples.

Alias table An alias table contains entity candidates for a given mention string and limits the possibilities to a relatively small set. Such tables are often compiled from a labeled training set and domain-specific heuristics.

Frequency statistics Many systems use frequency statistics obtained from a large labeled corpus to estimate entity popularity and the probability of a mention string linking to an entity. These statistics are very powerful when available.

Structured data Some systems assume access to structured data such as relationship tuples (e.g., (*Barack Obama*, *Spouse*, *Michelle Obama*)) or a type hierarchy to aid disambiguation.

2.2 Task Definition

The main motivation for this task is to expand the scope of entity linking systems and make them generalizable to unseen entity sets for which none of the powerful resources listed above are readily available. Therefore, we drop the above assumptions and make one weak assumption: the existence of an **entity dictionary** $\mathcal{E} = \{(e_i, d_i)\}_{i=1,..,K}$, where d_i is a text description of entity e_i.

Our goal is to build entity linking systems that can generalize to new domains and entity dictionaries, which we term *worlds*. We define a world as $\mathcal{W} = (\mathcal{M}_\mathcal{W}, \mathcal{U}_\mathcal{W}, \mathcal{E}_\mathcal{W})$, where $\mathcal{M}_\mathcal{W}$ and $\mathcal{U}_\mathcal{W}$ are distributions over mentions and documents from the world, respectively, and $\mathcal{E}_\mathcal{W}$ is an entity dictionary associated with $\mathcal{W}$. Mentions m from $\mathcal{M}_\mathcal{W}$

Task	In-Domain	Seen Entity Set	Small Candidate Set	Statistics	Structured Data	Entity dictionary
Standard EL	✓	✓		✓	✓	✓
Cross-Domain EL		✓		✓	✓	✓
Linking to Any DB (Sil et al., 2012)			✓		✓	✓
Zero-Shot EL						✓

Table 1: Assumptions and resources for entity linking task definitions. We classify task definitions based on whether (i) the system is tested on mentions from the training domain (In-Domain), (ii) linked mentions from the target entity set are seen during training (Seen Entity Set), (iii) a small high-coverage candidate set can be derived using alias tables or strict token overlap constraints (Small Candidate Set) and the availability of (iv) Frequency statistics, (v) Structured Data, and (vi) textual descriptions (Entity dictionary).

are defined as mention spans in documents from $\mathcal{U}_{\mathcal{W}}$. We assume the availability of labelled mention, entity pairs from one or more source worlds $\mathcal{W}^1_{\text{src}} \dots \mathcal{W}^n_{\text{src}}$ for training. At test time we need to be able to label mentions in a new world $\mathcal{W}_{\text{tgt}}$. Note that the entity sets $\mathcal{E}_{\mathcal{W}^1_{\text{src}}}, \dots, \mathcal{E}_{\mathcal{W}^n_{\text{src}}}, \mathcal{E}_{\mathcal{W}_{\text{tgt}}}$ are disjoint. See Figure 1 for an illustration of several training and test worlds.

We additionally assume that samples from the document distribution $\mathcal{U}_{\mathcal{W}_{\text{tgt}}}$ and the entity descriptions $\mathcal{E}_{\mathcal{W}_{\text{tgt}}}$ are available for training. These samples can be used for unsupervised adaptation to the target world. During training, mention boundaries for mentions in $\mathcal{W}_{\text{tgt}}$ are not available. At test time, mention boundaries are provided as input.

2.3 Relationship to other EL tasks

We summarize the relationship between the newly introduced zero-shot entity linking task and prior EL task definitions in Table 1.

Standard EL While there are numerous differences between EL datasets (Bunescu and Pasca, 2006; Ling et al., 2015), most focus on a standard setting where mentions from a comprehensive test entity dictionary (often Wikipedia) are seen during training, and rich statistics and meta-data can be utilized (Roth et al., 2014). Labeled in-domain documents with mentions are also assumed to be available.

Cross-Domain EL Recent work has also generalized to a cross-domain setting, linking entity mentions in different types of text, such as blogposts and news articles to the Wikipedia KB, while only using labeled mentions in Wikipedia for training (e.g., Gupta et al. (2017); Le and Titov (2018), *inter alia*).

Linking to Any DB Sil et al. (2012) proposed a task setup very similar to ours, and later work (Wang et al., 2015) has followed a similar setting. The main difference between zero-shot EL and these works is that they assumed either a high-coverage alias table or high-precision token overlap heuristics to reduce the size of the entity candidate set (i.e., to less than four in Sil et al. (2012)) and relied on structured data to help disambiguation. By compiling and releasing a multi-world dataset focused on learning from textual information, we hope to help drive progress in linking entities for a broader set of applications.

Work on word sense disambiguation based on dictionary definitions of words is related as well (Chaplot and Salakhutdinov, 2018), but this task exhibits lower ambiguity and existing formulations have not focused on domain generalization.

3 Dataset Construction

We construct a new dataset to study the zero-shot entity linking problem using documents from Wikia.[3] Wikias are community-written encyclopedias, each specializing in a particular subject or theme such as a fictional universe from a book or film series. Wikias have many interesting properties suitable for our task. Labeled mentions can be automatically extracted based on hyperlinks. Mentions and entities have rich document context that can be exploited by reading comprehension approaches. Each Wikia has a large number of unique entities relevant to a specific theme, making it a useful benchmark for evaluating domain generalization of entity linking systems.

We use data from 16 Wikias, and use 8 of them for training and 4 each for validation and testing. To construct data for training and evaluation, we first extract a large number of mentions from the Wikias. Many of these mentions can be easily linked by string matching between mention string

[3] https://www.wikia.com.

World	Entities	Mentions		
		Train	Evaluation	
			Seen	Unseen
Training				
American Football	31929	3898	410	333
Doctor Who	40281	8334	819	702
Fallout	16992	3286	337	256
Final Fantasy	14044	6041	629	527
Military	104520	13063	1356	1408
Pro Wrestling	10133	1392	151	111
StarWars	87056	11824	1143	1563
World of Warcraft	27677	1437	155	100
Validation				
Coronation Street	17809	0	0	1464
Muppets	21344	0	0	2028
Ice Hockey	28684	0	0	2233
Elder Scrolls	21712	0	0	4275
Test				
Forgotten Realms	15603	0	0	1200
Lego	10076	0	0	1199
Star Trek	34430	0	0	4227
YuGiOh	10031	0	0	3374

Table 2: Zero-shot entity linking dataset based on Wikia.

Coronation Street		
Mention	She told ray that Dickie and Audrey had met up again and tried to give their marriage another go . . . I don't want to see **her** face again . . . "	
	Dickie Fleming	Richard "Dickie" Fleming lived in coronation street with his wife Audrey from 1968 to 1970.
✓	Audrey Fleming	Audrey Fleming (neé bright) was a resident of 3 coronation street from 1968 to 1970 . Audrey married Dickie Fleming . . .
	Zeedan Nazir	Zeedan Nazir is the son of the Late Kal and Jamila Nazir . . .
Star Wars		
Mention	The droid acted as Moff Kilran's representative on board the Black Talon, an **Imperial transport ship**.	
✓	Gage-class transport	The Gage-class transport was a transport design used by the reconstituted Sith Empire of the Great Galactic War.
	Imperial Armored Transport	The Kuat Drive Yards Imperial Armored Transport was fifty meters long and carried ten crewmen and twenty soldiers.
	M-class Imperial Attack Transport	The M-class Imperial Attack Transport was a type of starship which saw service in the Imperial Military during the Galactic War.

Table 3: Example mention and entity candidates from Coronation Street and Star Wars. Note that the language usage is very different across different Worlds.

and the title of entity documents. These mentions are downsampled during dataset construction, and occupy a small percentage (5%) of the final dataset. While not completely representative of the natural distribution of mentions, this data construction method follows recent work that focuses on evaluating performance on the challenging aspects of the entity linking problem (e.g., Gupta et al. (2017) selected mentions with multiple possible entity candidates for assessing in-domain unseen entity performance). Each Wikia document corresponds to an entity, represented by the title and contents of the document. These entities, paired with their text descriptions, comprise the entity dictionary.

Since the task is already quite challenging, we assume that the target entity exists in the entity dictionary and leave NIL recognition or clustering (NIL mentions/entities refer to entities non-existent in the knowledge-base) to future editions of the task and dataset.

We categorize the mentions based on token overlap between mentions and the corresponding entity title as follows. *High Overlap*: title is identical to mention text, *Multiple Categories*: title is mention text followed by a disambiguation phrase (e.g., mention string: 'Batman', title: 'Batman (Lego)'), *Ambiguous substring*: mention is a substring of title (e.g., mention string: 'Agent', title: 'The Agent'). All other mentions are categorized as *Low Overlap*. These mentions respectively constitute approximately 5%, 28%, 8% and 59% of the mentions in the dataset.

Table 2 shows some statistics of the dataset. Each domain has a large number of entities ranging from 10,000 to 100,000. The training set has 49,275 labeled mentions. To examine the in-domain generalization performance, we construct heldout sets *seen* and *unseen* of 5,000 mentions each, composed of mentions that link to only entities that were seen or unseen during training, respectively. The validation and test sets have 10,000 mentions each (all of which are unseen).

Table 3 shows examples of mentions and entities in the dataset. The vocabulary and language used in mentions and entity descriptions differs drastically between the different domains. In addition to acquiring domain specific knowledge, understanding entity descriptions and performing reasoning is required in order to resolve mentions.

4 Models for Entity Linking

We adopt a two-stage pipeline consisting of a fast candidate generation stage, followed by a more expensive but powerful candidate ranking stage.

4.1 Candidate generation

Without alias tables for standard entity linking, a natural substitute is to use an IR approach for candidate generation. We use BM25, a variant of TF-IDF to measure similarity between mention string and candidate documents.[4] Top-k entities retrieved by BM25 scoring with Lucene[5] are used for training and evaluation. In our experiments k is set to 64. The coverage of the top-64 candidates is less than 77% on average, indicating the difficulty of the task and leaving substantial room for improvement in the candidate generation phase.

4.2 Candidate ranking

Since comparing two texts—a mention in context and a candidate entity description—is a task similar to reading comprehension and natural language inference tasks, we use an architecture based on a deep Transformer (Vaswani et al., 2017) which has achieved state-of-the-art performance on such tasks (Radford et al., 2018; Devlin et al., 2019).

As in BERT (Devlin et al., 2019), the mention in context m and candidate entity description e, each represented by 128 word-piece tokens, are concatenated and input to the model as a sequence pair together with special start and separator tokens: ($[\texttt{CLS}]\ m\ [\texttt{SEP}]\ e\ [\texttt{SEP}]$). Mention words are signaled by a special embedding vector that is added to the mention word embeddings. The Transformer encoder produces a vector representation $h_{m,e}$ of the input pair, which is the output of the last hidden layer at the special pooling token $[\texttt{CLS}]$. Entities in a given candidate set are scored as $w^\top h_{m,e}$ where w is a learned parameter vector, and the model is trained using a softmax loss. An architecture with 12 layers, hidden dimension size 768 and 12 attention heads was used in our experiments. We refer to this model as **Full-Transformer**. By jointly encoding the entity description and the mention in context with a Transformer, they can attend to each other at every layer.

Note that prior neural approaches for entity linking have not explored such architectures with deep cross-attention. To assess the value of this departure from prior work, we implement the following two variants: (*i*) **Pool-Transformer**: a siamese-like network which uses two deep Transformers to separately derive single-vector representations of the mention in context, h_m, and the candidate entity, h_e; they take as input the mention in context and entity description respectively, together with special tokens indicating the boundaries of the texts: ($[\texttt{CLS}]\ m\ [\texttt{SEP}]$) and ($[\texttt{CLS}]\ e\ [\texttt{SEP}]$), and output the last hidden layer encoding at the special start token. The scoring function is $h_m^\top h_e$. Single vector representations for the two components have been used in many prior works, e.g., Gupta et al. (2017). (*ii*) **Cand-Pool-Transformer**: a variant which uses single vector entity representations but can attend to individual tokens of the mention and its context as in Ganea and Hofmann (2017). This architecture also uses two Transformer encoders, but introduces an additional attention module which allows h_e to attend to individual token representations of the mention in context.

In the experiments section, we also compare to re-implementations of Gupta et al. (2017) and Ganea and Hofmann (2017), which are similar to Pool-Transformer and Cand-Pool-Transformer respectively but with different neural architectures for encoding.

5 Adapting to the Target World

We focus on using unsupervised pre-training to ensure that downstream models are robust to target domain data. There exist two general strategies for pre-training: (1) task-adaptive pre-training, and (2) open-corpus pre-training. We describe these below, and also propose a new strategy: domain-adaptive pre-training (DAP), which is complementary to the two existing approaches.

Task-adaptive pre-training Glorot et al. (2011); Chen et al. (2012); Yang and Eisenstein (2015), *inter alia*, pre-trained on the source and target domain unlabeled data jointly with the goal of discovering features that generalize across domains. After pre-training, the model is fine-tuned on the source-domain labeled data.[6]

Open-corpus pre-training Instead of explicitly adapting to a target domain, this approach simply applies unsupervised pre-training to large corpora before fine-tuning on the source-domain labeled data. Examples of this approach include ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2019).

[4] We also experimented with using the mention+context text but this variant performs substantially worse.

[5] http://lucene.apache.org/

[6] In many works, the learned representations are kept fixed and only higher layers are updated.

Intuitively, the target-domain distribution is likely to be partially captured by pre-training if the open corpus is sufficiently large and diverse. Indeed, open-corpus pre-training has been shown to benefit out-of-domain performance far more than in-domain performance (He et al., 2018).

Domain-adaptive pre-training In addition to pre-training stages from other approaches, we propose to insert a penultimate *domain adaptive pre-training* (DAP) stage, where the model is pre-trained *only* on the target-domain data. As usual, DAP is followed by a final fine-tuning stage on the source-domain labeled data. The intuition for DAP is that representational capacity is limited, so models should prioritize the quality of target domain representations above all else.

We introduce notation to describe various ways in which pre-training stages can be composed.

- U_{src} denotes text segments from the union of source world document distributions $\mathcal{U}_{\mathcal{W}^1_{\text{src}}} \ldots \mathcal{U}_{\mathcal{W}^n_{\text{src}}}$.
- U_{tgt} denotes text segments from the document distribution of a target world $\mathcal{W}_{\text{tgt}}$.
- $U_{\text{src+tgt}}$ denotes randomly interleaved text segments from both U_{src} and U_{tgt}.
- U_{WB} denotes text segments from open corpora, which in our experiments are Wikipedia and the BookCorpus datasets used in BERT.

We can chain together a series of pre-training stages. For example, $U_{\text{WB}} \rightarrow U_{\text{src+tgt}} \rightarrow U_{\text{tgt}}$ indicates that the model is first pre-trained on the open corpus, then pre-trained on the combined source and target domains, then pre-trained on only the target domain, and finally fine-tuned on the source-domain labeled data.[7] We show that chaining together different pre-training strategies provides additive gains.

6 Experiments

Pre-training We use the BERT-Base model architecture in all our experiments. The Masked LM objective (Devlin et al., 2019) is used for unsupervised pre-training. For fine-tuning language models (in the case of multi-stage pre-training) and fine-tuning on the Entity-Linking task, we use a small learning rate of 2e-5, following the recommendations from Devlin et al. (2019). For models trained from scratch we use a learning rate of 1e-4.

[7] We use the notation U_x interchangeably to mean both the unsupervised data x and the strategy to pre-train on x.

Model	Resources	Avg Acc
Edit-distance	$\emptyset$	16.49
TF-IDF [8]	$\emptyset$	26.06
Ganea and Hofmann (2017)	GloVe	26.96
Gupta et al. (2017)	GloVe	27.03
Full-Transformer	$\emptyset$	19.17
Full-Transformer (Pre-trained)	U_{src}	66.55
Full-Transformer (Pre-trained)	U_{tgt}	67.87
Full-Transformer (Pre-trained)	$U_{\text{src+tgt}}$	67.91
Pool-Transformer (Pre-trained)	U_{WB}	57.61
Cand-Pool-Trans. (Pre-trained)	U_{WB}	52.62
Full-Transformer (Pre-trained)	U_{WB}	76.06

Table 4: Baseline results for Zero-shot Entity Linking. Averaged normalized Entity-Linking accuracy on all validation domains. $U_{\text{src+tgt}}$ refers to masked language model pre-training on unlabeled data from training and validation worlds.

Evaluation We define the *normalized* entity-linking performance as the performance evaluated on the subset of test instances for which the gold entity is among the top-k candidates retrieved during candidate generation. The *unnormalized* performance is computed on the entire test set. Our IR-based candidate generation has a top-64 recall of 76% and 68% on the validation and test sets, respectively. The unnormalized performance is thus upper-bounded by these numbers. Strengthening the candidate generation stage improves the unnormalized performance, but this is outside the scope of our work. Average performance across a set of worlds is computed by macro-averaging. Performance is defined as the accuracy of the single-best identified entity (top-1 accuracy).

6.1 Baselines

We first examine some baselines for zero-shot entity linking in Table 4. We include naive baselines such as Levenshtein edit-distance and TF-IDF, which compare the mention string against candidate entity title and full document description, respectively, to rank candidate entities.

We re-implemented recent neural models designed for entity linking (Ganea and Hofmann, 2017; Gupta et al., 2017), but did not expect them to perform well since the original systems were designed for settings where labeled mentions or meta-data for the target entities were available.

(a)

Pretraining	$\mathcal{W}^1_{\text{tgt}}$	$\mathcal{W}^2_{\text{tgt}}$	$\mathcal{W}^3_{\text{tgt}}$	$\mathcal{W}^4_{\text{tgt}}$	Avg
$U_{\text{src+tgt}}$ (Glorot et al., 2011)†	73.19	71.61	62.16	64.69	67.91
$U_{\text{src+tgt}} \rightarrow U_{\text{tgt}}$ (DAP)	79.20	75.55	66.85	66.72	**72.08**
U_{WB} (Devlin et al., 2019)	83.40	79.00	73.03	68.82	76.06
$U_{\text{WB}} \rightarrow U_{\text{tgt}}$ (DAP)	81.68	81.34	73.17	71.97	**77.04**
$U_{\text{WB}} \rightarrow U_{\text{src+tgt}}$	82.92	79.00	72.62	69.55	76.02
$U_{\text{WB}} \rightarrow U_{\text{src+tgt}} \rightarrow U_{\text{tgt}}$ (DAP)	82.82	81.59	75.34	72.52	**78.07**

(b)

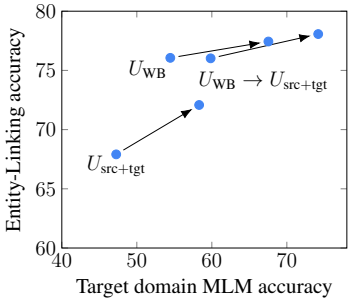

Figure 2: **Left: (a)** Impact of using Domain Adaptive Pre-training. We fine-tune all the models on the source labeled data after pretraining. **Right: (b)** Relationship between MLM (Masked LM) accuracy of pre-trained model and Entity-Linking performance of the fine-tuned model, evaluated on target domains. Adding domain adaptive pre-training improves both MLM accuracy as well as the entity linking performance. **Note:** *src* represents the union of all 8 training worlds and we adapt to one *tgt* world at a time. The target worlds are $\mathcal{W}^1_{\text{tgt}}$: *Coronation street*, $\mathcal{W}^2_{\text{tgt}}$: *Muppets*, $\mathcal{W}^3_{\text{tgt}}$: *Ice hockey*, $\mathcal{W}^4_{\text{tgt}}$: *Elder scrolls*. †We refer to Glorot et al. (2011) for the idea of training a denoising autoencoder on source and target data together rather than the actual implementation. See text for more details.

The poor performance of these models validates the necessity of using strong reading comprehension models for zero-shot entity linking.

When using the Full-Transformer model, pre-training is necessary to achieve reasonable performance. We present results for models pre-trained on different subsets of our task corpus (U_{src}, U_{tgt}, $U_{\text{src+tgt}}$) as well as pre-training on an external large corpus (U_{WB}). We observe that the choice of data used for pre-training is important.

In Table 4 we also compare the Pool-Transformer, Candidate-Pool-Transformer and Full-Transformer. The significant gap between Full-Transformer and the other variants shows the importance of allowing fine-grained comparisons between the two inputs via the cross attention mechanism embedded in the Transformer. We hypothesize that prior entity linking systems did not need such powerful reading comprehension models due to the availability of strong additional meta information. The remaining experiments in the paper use the Full-Transformer model, unless mentioned otherwise.

6.2 Generalization to Unseen Entities and New Worlds

To analyze the impact of unseen entities and domain shift in zero-shot entity linking, we evaluate performance on a more standard in-domain entity linking setting by making predictions on held out mentions from the training worlds. Table 5 compares entity linking performance for different entity splits. Seen entities from the training worlds are unsurprisingly the easiest to link to. For unseen entities from the training world, we observe a 5-point drop in performance. Entities from new worlds (which are by definition unseen and are mentioned in out-of-domain text) prove to be the most difficult. Due to the shift in both the language distribution and entity sets, we observe a 11-point drop in performance. This large generalization gap demonstrates the importance of adaptation to new worlds.

Evaluation	Accuracy
Training worlds, seen	87.74
Training worlds, unseen	82.96
Validation worlds, unseen	76.06

Table 5: Performance of the Full-Transformer (U_{WB}) model evaluated on seen and unseen entities from the training and validation worlds.

6.3 Impact of Domain Adaptive Pre-training

Our experiments demonstrate that DAP improves on three state-of-the-art pre-training strategies:

- $U_{\textbf{src+tgt}}$: task-adaptive pre-training, which combines source and target data for pre-training (Glorot et al., 2011).[9]
- $U_{\textbf{WB}}$: open-corpus pre-training, which uses Wikipedia and the BookCorpus for pre-training (We use a pre-trained BERT model (Devlin et al., 2019)).
- $U_{\textbf{WB}} \rightarrow U_{\textbf{src+tgt}}$: the previous two strategies chained together. While no prior work has applied this approach to domain adaptation, a similar approach for task adaptation was proposed by Howard and Ruder (2018).

[9]We use Masked LM and Transformer encoder, which are more powerful than the instantiation in (Glorot et al., 2011).

Pre-training	EL Accuracy N. Acc.	U. Acc.
U_{WB} (Devlin et al., 2019)	75.06	55.08
$U_{\text{WB}} \rightarrow U_{\text{tgt}}$ (DAP)	76.17	55.88
$U_{\text{WB}} \rightarrow U_{\text{src+tgt}} \rightarrow U_{\text{tgt}}$ (DAP)	**77.05**	**56.58**

Table 6: Performance on test domains with Full-Transformer. **N. Acc** represents the normalized accuracy. **U. Acc** represents the unnormalized accuracy. The unnormalized accuracy is upper-bounded by 68%, the top-64 recall of the candidate generation stage.

Mention Category	Recall@64	EL Accuracy N. Acc.	U. Acc.
High Overlap	99.28	87.64	87.00
Ambiguous Substring	88.03	75.89	66.81
Multiple categories	84.88	77.27	65.59
Low Overlap	54.37	71.46	38.85

Table 7: Performance on test domains categorized by mention categories. Recall@64 indicates top-64 performance of candidate generation. N. Acc. and U. Acc. are respectively the normalized and unnormalized accuracies.

The results are in Figure 2(a). DAP improves all pre-training strategies with an additional pre-training stage on only target-domain data. The best setting, $U_{\text{WB}} \rightarrow U_{\text{src+tgt}} \rightarrow U_{\text{tgt}}$, chains together all existing strategies. DAP improves the performance over a strong pre-trained model (Devlin et al., 2019) by 2%.

To further analyze the results of DAP, we plot the relationships between the accuracy of Masked LM (MLM accuracy) on target unlabeled data and the final target normalized accuracy (after fine-tuning on the source labeled data) in Figure 2(b). Adding an additional pre-training stage on the target unlabeled data unsurprisingly improves the MLM accuracy. More interestingly, we find that improvements in MLM accuracy are consistently followed by improvements in entity linking accuracy. It is intuitive that performance on unsupervised objectives reflect the quality of learned representations and correlate well with downstream performance. We show empirically that this trend holds for a variety of pre-training strategies.

6.4 Test results and performance analysis

Table 6 shows the normalized and unnormalized Entity Linking performance on test worlds. Our best model that chains together all pre-training strategies achieves normalized accuracy of 77.05% and unnormalized accuracy of 56.58%. Note that the unnormalized accuracy corresponds to identifying the correct entity from tens of thousands of candidate entities.

To analyze the mistakes made by the model, we compare EL accuracy across different mention categories in Table 7. Candidate generation (Recall@64) is poor in the Low Overlap category. However, the ranking model performs in par with other hard categories for these mentions. Overall EL accuracy can thus be improved significantly by strengthening candidate generation.

7 Related Work

We discussed prior entity linking task definitions and compared them to our task in section 2. Here, we briefly overview related entity linking models and unsupervised domain adaptation methods.

Entity linking models Entity linking given mention boundaries as input can be broken into the tasks of candidate generation and candidate ranking. When frequency information or alias tables are unavailable, prior work has used measures of similarity of the mention string to entity names for candidate generation (Sil et al., 2012; Murty et al., 2018). For candidate ranking, recent work employed distributed representations of mentions in context and entity candidates and neural models to score their compatibility. Mentions in context have been represented using e.g., CNN (Murty et al., 2018), LSTM (Gupta et al., 2017), or bag-of-word embeddings (Ganea and Hofmann, 2017). Entity descriptions have been represented using similar architectures. To the best of our knowledge, while some models allow for cross-attention between single-vector entity embeddings and mention-in-context token representations, no prior works have used full cross-attention between mention+context and entity descriptions.

Prior work on entity linking tasks most similar to ours used a linear model comparing a mention in context to an entity description and associated structured data (Sil et al., 2012). Sil et al. (2012) also proposed a distant supervision approach which could use first-pass predictions for mentions in the target domain as noisy supervision for re-training an in-domain model. We believe this approach is complementary to unsupervised representation learning and could bring additional benefits. In another task similar to ours, Wang et al. (2015) used collective inference and

target database relations to obtain good performance without (domain, target database)-specific labeled training data. Collective inference is another promising direction, but could have limited success when no metadata is available.

Unsupervised domain adaptation There is a large body of work on methods for unsupervised domain adaptation, where a labeled training set is available for a source domain and unlabeled data is available for the target domain. The majority of work in this direction assume that training and test examples consist of (x, y) pairs, where y is in a fixed shared label set $\mathcal{Y}$. This assumption holds for classification and sequence labeling, but not for zero-shot entity linking, since the source and target domains have disjoint labels.

Most state-of-the-art methods learn non-linear shared representations of source and target domain instances, through denoising training objectives (Eisenstein, 2018). In Section 5, we overviewed such work and proposed an improved domain adaptive pre-training method.

Adversarial training methods (Ganin et al., 2016), which have also been applied to tasks where the space $\mathcal{Y}$ is not shared between source and target domains (Cohen et al., 2018), and multi-source domain adaptation methods (Zhao et al., 2018; Guo et al., 2018) are complementary to our work and can contribute to higher performance.

8 Conclusion

We introduce a new task for zero-shot entity linking, and construct a multi-world dataset for it. The dataset can be used as a shared benchmark for entity linking research focused on specialized domains where labeled mentions are not available, and entities are defined through descriptions alone. A strong baseline is proposed by combining powerful neural reading comprehension with domain-adaptive pre-training.

Future variations of the task could incorporate NIL recognition and mention detection (instead of mention boundaries being provided). The candidate generation phase leaves significant room for improvement. We also expect models that jointly resolve mentions in a document would perform better than resolving them in isolation.

Acknowledgements

We thank Rahul Gupta and William Cohen for providing detailed helpful feedback on an earlier draft of this paper. We thank the Google AI Language Team for valuable suggestions and feedback.

References

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*.

Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *The 41st International ACM SIGIR Conference on Research; Development in Information Retrieval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jacob Eisenstein. 2018. *Natural Language Processing*. MIT Press.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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*.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. *arXiv preprint arXiv:1805.04787*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

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

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Dan Roth, Heng Ji, Ming-Wei Chang, and Taylor Cassidy. 2014. Wikification and beyond: The challenges of entity and concept grounding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*.

Avi Sil, Heng Ji, Dan Roth, and Silviu-Petru Cucerzan. 2018. Multi-lingual entity discovery and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Tutorial Abstracts*.

Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking named entities to any database. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Han Wang, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*.

A Examining model errors and predictions

In tables 8, 9, 10, 11 we show some example mentions and model predictions. For each instance, the examples show the correct gold entity and the top-5 predictions from the model. Examples show 32 token contexts centered around mentions and the first 32 tokens of candidate entity documents.

	Coronation Street	
Mention	Robbie pulled over the ambulance with a van and used a gun to get the **Prison Officer** with Tony to release him . He integrated himself with the Street residents , finding	
Gold Entity	Prison Officer (Episode 7351)	The unnamed Prison Officer was on duty during May 2010 in the Highfield Prison dining room when Tony Gordon provoked a fight with a fellow inmate
	Top-5 predictions	
✓	Prison Officer (Episode 7351)	The unnamed Prison Officer was on duty during May 2010 in the Highfield Prison dining room when Tony Gordon provoked a fight with a fellow inmate
	Inmate (Episode 7351)	The Inmate was an unnamed fellow prisoner of Tony Gordon in Highfield Prison . Tony provoked a fight in the dining room with the inmate by staring
	Police Officer (Simon Willmont)	The unnamed Police Officer was on duty at Weatherfield Police Station in March 2010 when Peter Barlow was released from custody following his arrest as he
	Prison Officer (Bill Armstrong)	The Prison Officer looked after the incarceration of three Coronation Street residents : In November 2000 he was on duty at Strangeways Jail when Jim McDonald
	Robbie Sloane	Quietly spoken Robbie Sloane was Tony Gordon ' s henchman and a convicted murderer , who he met while sharing a cell at Highfield Prison in 2010 . When Robbie

Table 8: Mention and entity candidates from Coronation Street.

	Muppets	
Mention	Bean Bunny was introduced during the seventh season of " Muppet Babies " , and a **pre - teen Bean** would later be featured as part of the Muppet Kids series . Bean was active	
Gold Entity	Bean Bunny (Muppet Kids)	A young version of Bean Bunny made a few appearances in the Muppet Kids books and video games . Young Bean moves to the Muppet Kids
	Top-5 predictions	
	Baby Bean Bunny	Baby Bean Bunny appeared in the late 1989 / 1990 seasons of " Muppet Babies " as a baby version of Bean Bunny . He joined the other babies
✓	Bean Bunny (Muppet Kids)	A young version of Bean Bunny made a few appearances in the Muppet Kids books and video games . Young Bean moves to the Muppet Kids
	Bean Bunny	Bean Bunny first appeared in 1986 as the star of the TV special " The Tale of the Bunny Picnic " . The cute bunny was part of a family
	Piggy (Muppet Kids)	A pre - teen version of Miss Piggy , as seen in the " Muppet Kids " books and video games . Piggy lives in a fancy
	Muppet Kids	Muppet Kids was a series of books and educational software made in the 1990s , featuring young , pre - teen versions of the principal franchise characters . Characters included

Table 9: Mention and entity candidates from Muppets.

	Ice Hockey	
Mention	1979 - 80 PCJHL Season This is a list of **Peace - Cariboo Junior Hockey League** Standings for the 1979 - 80 season . This was the PCJHL ' s final	
Gold Entity	Rocky Mountain Junior Hockey League	The Rocky Mountain Junior Hockey League was a Canadian Junior " A " ice hockey league in British Columbia . History . Promoted to a Junior "
	Top-5 predictions	
	Peace Junior Hockey League	Hockey League Peace Junior Hockey League is a League that started in the 1960 ' s and ended in 1975 . Then change its name to Peace Cariboo junior Hockey
	Cariboo Hockey League	The Cariboo Hockey League was a Senior and Intermediate hockey league in the Cariboo District of British Columbia , Canada . History . The league began in the 1955
	Cariboo Junior League	The Cariboo Junior League operated in northern British Columbia in the 1963 - 64 season . Its champion was eligible for the British Columbia Junior Playoffs . The league
✓	Rocky Mountain Junior Hockey League	The Rocky Mountain Junior Hockey League was a Canadian Junior " A " ice hockey league in British Columbia . History . Promoted to a Junior "
	North West Junior Hockey League	The North West Junior Hockey League is a Junior " B " ice hockey league operating in the Peace River region of Alberta and British Columbia ,

Table 10: Mention and entity candidates from Ice Hockey.

	Elder Scrolls	
Mention	to get everyone to safety . Rolunda ' s brother is one of those people . **The Frozen Man** . Rolunda ' s brother Eiman has ventured into Orkey ' s Hollow to find	
Gold Entity	The Frozen Man (Quest)	The Frozen Man is a quest available in The Elder Scrolls Online. It involves finding a Nord who has been trapped in ice by a mysterious " Frozen Man
	Top-5 predictions	
✓	The Frozen Man (Quest)	The Frozen Man is a quest available in The Elder Scrolls Online. It involves finding a Nord who has been trapped in ice by a mysterious " Frozen Man
	The Frozen Man	The Frozen Man is an insane Bosmer ghost found in Orkey ' s Hollow . He says he was in a group of people inside the cave when it
	Kewan	Kewan is a Redguard worshipper of the Daedric Prince Peryite . He is frozen in a trance that relates to the Daedric quest , but can be unfrozen in completion the
	Stromgruf the Steady	Stromgruf the Steady is the Nord who is found in the Grazelands of Vvardenfell , west of Pulk and east of Vassamsi Grotto (Online) . He is
	Maren the Seal	Maren the Seal is a Nord hunter and worshipper of the Daedric Prince Peryite . She is frozen in a trance that relates to the Daedric Prince ' s

Table 11: Mention and entity candidates from Elder Scrolls.