EntEval: A Holistic Evaluation Benchmark for Entity Representations

Mingda Chen[3*] **Zewei Chu**[1*] **Yang Chen**[2] **Karl Stratos**[4†] **Kevin Gimpel**[3]
[1]University of Chicago, IL, USA
[2]Ohio State University, OH, USA
[3]Toyota Technological Institute at Chicago, IL, USA
[4]Rutgers University, NJ, USA
{mchen,kgimpel}@ttic.edu,stratos@cs.rutgers.edu,zeweichu@uchicago.edu,chen.9279@osu.edu

Abstract

Rich entity representations are useful for a wide class of problems involving entities. Despite their importance, there is no standardized benchmark that evaluates the overall quality of entity representations. In this work, we propose EntEval: a test suite of diverse tasks that require nontrivial understanding of entities including entity typing, entity similarity, entity relation prediction, and entity disambiguation. In addition, we develop training techniques for learning better entity representations by using natural hyperlink annotations in Wikipedia. We identify effective objectives for incorporating the contextual information in hyperlinks into state-of-the-art pretrained language models (Peters et al., 2018a) and show that they improve strong baselines on multiple EntEval tasks.[1]

1 Introduction

Entity representations play a key role in numerous important problems including language modeling (Ji et al., 2017), dialogue generation (He et al., 2017), entity linking (Gupta et al., 2017), and story generation (Clark et al., 2018). One successful line of work on learning entity representations has been learning *static* embeddings: that is, assign a unique vector to each entity in the training data (Gupta et al., 2017; Yamada et al., 2016, 2017). While these embeddings are useful in many applications, they have the obvious drawback of not accommodating unknown entities. Another limiting factor is the lack of an evaluation benchmark: it is often difficult to know which entity representations are better for which tasks.

We introduce EntEval: a carefully designed benchmark for holistically evaluating entity representations. It is a test suite of diverse tasks that require nontrivial understanding of entities, including entity typing, entity similarity, entity relation prediction, and entity disambiguation. Motivated by the recent success of contextualized word representations (henceforth: CWRs) from pretrained models (McCann et al., 2017; Peters et al., 2018a; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019b), we propose to encode the mention context or the description to dynamically represent an entity. In addition, we perform an in-depth comparison of ELMo and BERT-based embeddings and find that they show different characteristics on different tasks. We analyze each layer of the CWRs and make the following observations:

- The dynamically encoded entity representations show a strong improvement on the entity disambiguation task compared to prior work using static entity embeddings.
- BERT-based entity representations require further supervised training to perform well on downstream tasks, while ELMo-based representations are more capable of performing zero-shot tasks.
- In general, higher layers of ELMo and BERT-based CWRs are more transferable to EntEval tasks.

To further improve contextualized and descriptive entity representations (CER/DER), we leverage natural hyperlink annotations in Wikipedia. We identify effective objectives for incorporating the contextual information in hyperlinks and improve ELMo-based CWRs on a variety of entity related tasks.

[*] Equal contribution. Listed in alphabetical order.

[†] Work done while the author was at Toyota Technological Institute at Chicago.

[1] Data processing and evaluation scripts are available at https://github.com/ZeweiChu/EntEval

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing*, pages 421–433,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

2 Related Work

EntEval and the training objectives considered in this work are built on previous works that involve reasoning over entities. We give a brief overview of relevant works.

Entity linking/disambiguation. Entity linking is a fundamental task in information extraction with a wealth of literature (He et al., 2013; Guo and Barbosa, 2014; Ling et al., 2015; Huang et al., 2015; Francis-Landau et al., 2016; Le and Titov, 2018; Martins et al., 2019). The goal of this task is to map a mention in context to the corresponding entity in a database. A natural approach is to learn entity representations that enable this mapping. Recent works focused on learning a fixed embedding for each entity using Wikipedia hyperlinks (Yamada et al., 2016; Ganea and Hofmann, 2017; Le and Titov, 2019). Gupta et al. (2017) additionally train context and description embeddings jointly, but this mainly aims to improve the quality of the fixed entity embeddings rather than using the context and description embeddings directly; we find that their context and description encoders perform poorly on EntEval tasks.

A closely related concurrent work by (Logeswaran et al., 2019) jointly encodes a mention in context and an entity description from Wikia to perform zero-shot entity linking. In contrast, here we seek to pretrain a general purpose entity representations that can function well either given or not given entity descriptions or mention contexts.

Other entity-related tasks involve entity typing (Yaghoobzadeh and Schütze, 2015; Murty et al., 2017; Del Corro et al., 2015; Rabinovich and Klein, 2017; Choi et al., 2018; Onoe and Durrett, 2019; Obeidat et al., 2019) and coreference resolution (Durrett and Klein, 2013; Wiseman et al., 2016; Lee et al., 2017; Webster et al., 2018; Kantor and Globerson, 2019).

Evaluating pretrained representations. Recent work has sought to evaluate the knowledge acquired by pretrained language models (Shi et al., 2016; Adi et al., 2017; Belinkov et al., 2017; Peters et al., 2018b; Conneau et al., 2018; Conneau and Kiela, 2018; Wang et al., 2018; Liu et al., 2019a; Chen et al., 2019a, *inter alia*). In this work, we focus on evaluating their capabilities in modeling entities.

Part of EntEval involves evaluating world knowledge about entities, relating them to fact checking (Vlachos and Riedel, 2014; Wang, 2017; Thorne et al., 2018; Yin and Roth, 2018; Chen et al., 2019b), and commonsense learning (Angeli and Manning, 2014; Bowman et al., 2015; Li et al., 2016; Mihaylov et al., 2018; Zellers et al., 2018; Trinh and Le, 2018; Talmor et al., 2019; Zellers et al., 2019; Sap et al., 2019; Rajani et al., 2019). Another related line of work is to integrate entity-related knowledge into the training of language models (Logan et al., 2019; Zhang et al., 2019; Sun et al., 2019).

Contextualized word representations. Contextualized word representations and pretrained language representation models, such as ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2018), are powerful pretrained models that have been shown to be effective for a variety of downstream tasks such as text classification, sentence relation prediction, named entity recognition, and question answering. Recent work has sought to evaluate the knowledge acquired by such models (Shi et al., 2016; Adi et al., 2017; Belinkov et al., 2017; Conneau et al., 2018; Conneau and Kiela, 2018; Liu et al., 2019a). In this work, we focus on evaluating their capabilities in modeling entities.

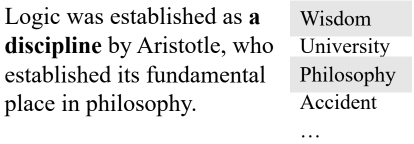
3 EntEval

We are interested in two approaches: contextualized entity representations (henceforth: CER) and descriptive entity representations (henceforth: DER), both encoding fixed-length vector representations for entities.

The contextualized entity representations encodes an entity based on the context it appears regardless of whether the entity is seen before. The motivation behind contextualized entity representations is that we want an entity encoder that does not depend on entries in a knowledge base, but is capable of inferring knowledge about an entity from the context it appears.

As opposed to contextualized entity representations, descriptive entity representations do rely on entries in Wikipedia. We use a model-specific function f to obtain a fixed-length vector representation from the entity's textual description.

To evaluate CERs and DERs, we propose a wide range of entity related tasks. Since our purpose is for examining the learned entity representations, we only use a linear classifier and freeze the entity representations when performing the follow-



Logic was established as **a discipline** by Aristotle, who established its fundamental place in philosophy.

Wisdom
University
Philosophy
Accident
...

Figure 1: An example taken from ET. Targeted entity mention is bold. Candidate categories are on the right. Gold standard categories are in gray.

ing tasks. Unless otherwise noted, when the task involves a pair of entities, the input to the classifier are the entity representations x_1 and x_2, concatenated with their element-wise product and absolute difference: $[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$. This input format has been used in SentEval (Conneau and Kiela, 2018).

The datasets used in EntEval tasks are summarized in table 1. It shows the number of instances in train/valid/test split for each dataset, and the number of target classes if this is a classification task. We describe the proposed tasks in the following subsections.

3.1 Entity Typing (ET)

The task of entity typing (ET) is to assign types to an entity given only the context of the entity mention. ET is context-sensitive, making it an effective approach to probe the knowledge of context encoded in pretrained representations. For example, in the sentence "Bill Gates has donated billions to eradicate malaria", "Bill Gates" has the type of "philanthropist" instead of "inventor" (Choi et al., 2018). In this task, we will contextualized entity representations, followed by a linear layer to make predictions. We use the annotated ultra-fine entity typing dataset of Choi et al. (2018) with standard data splits. As shown in Figure 1, there can be multiple labels for an instance. We use binary log loss for training using all positive and negative entity types, and report F_1 score. Thresholds are tuned based on validation set accuracy.

3.2 Coreference Arc Prediction (CAP)

Given two entities and the associated context, the task is to determine whether they refer to the same entity. Solving this task may require the knowledge of entities. For example, in the sentence "Revenues of \$14.5 billion were posted by Dell$_1$. The company$_1$ ...", there is no prior context of "Dell", so having known "Dell" is a company instead of the people "Michael Dell" will surely benefit the model (Durrett and Klein, 2014). Unlike other tasks, coreference typically involves longer context. To restrict the effect of broad context, we only keep two groups of coreference arcs from smaller context. One includes mentions that are in the same sentence ("same") for examining the model capability of encoding local context. The other includes mentions that are in consecutive sentences ("next") for the broader context. We create this task from the PreCo dataset (Chen et al., 2018), which has mentions annotated even when they are not part of coreference chains. We filter out instances in which both mentions are pronouns. All non-coreferent mention pairs are considered to be negative samples.

To make this task more challenging, for each instance we compute cosine similarity of mentions by averaging GloVe word vectors. We group the instances into bins by cosine similarity, and randomly select the same number of positive and negative instances from each bin to ensure that models do not solve this task by simply comparing similarity of mention names.

We use the contextualized entity representations of the two mentions to infer coreference arcs with supervised training and report the averaged accuracy of "same" and "next".

3.3 Entity Factuality Prediction (EFP)

The entity factuality prediction (EFP) task involves determining the correctness of statements regarding entities. We use the manually-annotated FEVER dataset (Thorne et al., 2018) for this task. FEVER is a task to verify whether a statement is supported by evidences. The original FEVER dataset includes three classes, namely "Supports", "Refutes", and "NotEnoughInfo" and evidences are additionally available for each instance. As our purpose is to examine the knowledge encoded in entity representations, we discard the last category ("NotEnoughInfo") and the evidence. In rare cases, instances in FEVER may include multiple entity mentions, so we randomly pick one. We randomly sample 10000, 2000, and 2000 instances for our training, validation, and test sets, respectively.

In this task, entity representations can be obtained either by contextualized entity representations or descriptive entity representations. In practice, we observe descriptive entity representations give better performance, which presumably is be-

	CAP		CERP	EFP	ET	KORE	WikiSRS		ERT	Rare	CoNLL
	same	next					Rel	Sim			
#train	3982	3982	10000	10000	1998	N/A	N/A	N/A	3130	10000	18538
#valid	3806	3828	2000	2000	1998	N/A	N/A	N/A	6260	4000	4790
#test	3938	3850	2000	2000	1998	20 × 20	688	688	6260	4000	4481
#classes	2		2	2	10331	N/A	N/A	N/A	626	4	up to 30

Table 1: Statistics of datasets used in EntEval tasks. CAP: coreference arc prediction, CERP: contexualized entity relationship prediction, EFP: entity factuality prediction, ET: entity typing, ESR: entity similarity and relatedness, ERT: entity relationship typing, NED: named entity disambiguation, Rare: rare entity prediction, CoNLL: CoNLL-YAGO named entity disambiguation.

REFUTES: The **New York City Landmarks Preservation Commission** consists of zero commissioners.
SUPPORTS: **TD Garden** has held Bruins games.

Figure 2: Two examples from the EFP.

TRUE: **Gin and vermouth** can **make a martini**
FALSE: **Connecticut** is not **a state**

Figure 3: Examples from the CERP.

Score	Entity Name
-	Apple Inc.
20	Steve Jobs
...	...
11	Microsoft
...	...
1	Ford Motor Company

Table 2: An example from KORE. The task is to rank the candidate entities by similarity.

cause these statements are more similar to descriptions than entity mentions. As shown in Figure 2, without providing additional evidences, solving this task requires knowledge of entities encoded in representations. We directly use entity representations as input to the classifier.

3.4 Contexualized Entity Relationship Prediction (CERP)

The task of contexualized entity relationship prediction (CERP) modeling determines the connection between two entities appeared in the same context. We use sentences from ConceptNet (Speer et al., 2017) with automatically parsed mentions and templates used to construct the dataset. We filter out non-English concepts and relations such as 'related', 'translation', 'synonym', and 'likely to find' since we seek to evaluate more complicated knowledge of entities encoded in representations. We further filter out non-entity mentions and entities with type 'DATE', 'TIME', 'PERCENT', 'MONEY', 'QUANTITY', 'ORDINAL', and 'CARDINAL' according to SpaCy (Honnibal and Montani, 2017). After filtering, we have 13374 assertions.

Negative samples are generated based on the following rules:

1. For each relationship, we replace an entity with similar negative entities based on cosine similarity of averaged GloVe embeddings (Pennington et al., 2014).
2. We change the relationship in positive samples from affirmation to negation (e.g., 'is' to 'is not'). These serve as negative samples.
3. We further sample positive samples from (1) in an attempt to prevent the 'not' token from being biased towards negative samples. Therefore, for negative samples we get from (1), we change the relationship from affirmation to negation as in (2) to get positive samples.

For example, let 'A is B' be the positive sample. (1) changes it to 'C is B' which serves as a negative sample and (2) changes it to 'A is not B' as another negative sample. (3) changes it to 'C is not B' as a positive example. In the end, we randomly sample 7000 instances from each class. This ends up yielding a 10000/2000/2000 train/dev/test dataset. As shown in Figure 3, this task cannot be solved by relying on surface form of sentences, instead it requires the input representations to encode knowledge of entities based on the context.

We use contextualized entity representations in this task.

3.5 Entity Similarity and Relatedness (ESR)

Given two entities with their descriptions from Wikipedia, the task is to determine their similarity or relatedness. After the entity descriptions are encoded into vector representations, we compute their cosine similarity as predictions. We use the KORE (Hoffart et al., 2012) and WikiSRS (Newman-Griffis et al., 2018) datasets in this task. Since the original datasets only provide entity names, we automatically add Wikipedia descriptions to each entity and manually ensure that every entity is matched to a Wikipedia description. We use Spearman's rank correlation coefficient between our computed cosine similarity and the gold standard similarity/relatedness scores to measure the performance of entity representations.

As KORE does not provide similarity scores of entity pairs, but simply ranks the candidate entities by their similarities to a target entity, we assign scores from 20 to 1 accordingly to each entity in the order of similarity. Table 2 shows an example from KORE. The fact that "Apple Inc." is more related to "Steve Jobs" than "Microsoft" requires multiple steps of inference, which motivates this task. Since the predictor we use is cosine similarity, which does not introduce additional parameters, we directly use encoded representations on the test set without any supervised training.

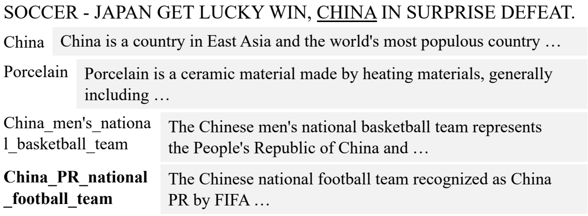
3.6 Entity Relationship Typing (ERT)

As another popular resource for common knowledge, we consider using Freebase (Bollacker et al., 2008) for probing the encoded knowledge by classifying the types of relations between pair of entities. First, we extract entity relation tuples (entity1, relation, entity2) from Freebase and then filter out easy tuples based on training a classifier using averaged GloVe vectors of entity names as input, which leaves us 626 types of relations, including "internet.website.owner", "film.film_art_director.films_art_directed", and "comic_books.comic_book_series.genre". We randomly sample 5 instances for each relation type to form our training set and 10 instances per type the for validation and test sets. We use Wikipedia descriptions for each entity in the pair whose relation we are predicting and we use descriptive entity representations for each entity with supervised training.

3.7 Named Entity Disambiguation (NED)

Named entity disambiguation is the task of linking a named-entity mention to its corresponding instance in a knowledge base such as Wikipedia. In this task, we consider CoNLL-YAGO (CoNLL; Hoffart et al., 2011) and Rare Entity Prediction (Rare; Long et al., 2017).

For CoNLL-YAGO, following Hoffart et al. (2011) and Yamada et al. (2016), we used the 27,816 mentions with valid entries in the knowledge base. For each entity mention m in its context, we generate a set of (at most) its top 30 candidate entities $C_m = \{c_j\}$ using CrossWikis (Spitkovsky and Chang, 2012). Some gold standard candidates c are not present in CrossWikis, so we set the prior probability $p_{\text{prior}}(y)$ for those to 1e-6 and normalize the resulting priors for the candidate entities. When adding Wikipedia descriptions, we manually ensure gold standard mentions are attached to a description, however, we discard candidate mentions that cannot be aligned to a Wikipedia page. We use contextualized entity representations for entity mentions and use descriptive entity representations for candidate entities. Training minimizes binary log loss using all negative examples. At test time, we use $\arg\max_{c\in C_m}[p_{\text{prior}}(c)+p_{\text{classifier}}(c)]$ as the prediction. We note that directly using prior as predictions yields an accuracy of 58.2%.



SOCCER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT.

China	China is a country in East Asia and the world's most populous country …
Porcelain	Porcelain is a ceramic material made by heating materials, generally including …
China_men's_national_basketball_team	The Chinese men's national basketball team represents the People's Republic of China and …
China_PR_national_football_team	The Chinese national football team recognized as China PR by FIFA …

Figure 4: An example from CoNLL-YAGO. Only four candidates are shown due to space constraints. The target mention is underlined. Sentences in gray are Wikipedia descriptions. The gold standard is boldfaced.

Long et al. (2017) introduce the task of *rare entity prediction*. The task has a similar format to CoNLL-YAGO entity linking. Given a document with a blank in it, the task is to select an entity from a provided list of entities with descriptions. Only rare entities are used in this dataset so that performing well on the task requires the ability to effectively represent entity descriptions. We randomly select 10k/4k/4k examples to construct train/valid/test sets. For simplicity, we only keep instances with four candidate entities.

Figure 4 shows an example from CoNLL-YAGO, where the "China" in context has many deceptive meanings. Here the candidate "China" has exact string match of the entity name but it should not be selected as it is an after-game report on soccer. To match the entities, this task requires both effective contextualize entity representations and descriptive entity representation.

Practically, we encode the context using CER to be x_1, and encode each entity description using DER to be x_2, and pass $[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$ to a linear model to predict whether it is the correct entity to fill in. The model is trained with cross entropy loss.

4 Methods

We first describe how we define encoders for contextualized entity representations (Section 4.1) and descriptive entity representations (Section 4.2), then we discuss how we train new encoders tailored to capture information from the hyperlink structure of Wikipedia (Section 4.3).

4.1 Encoders for Contextualized Entity Representations

For defining these encoders, we assume we have a sentence $s = (w_1, \ldots, w_T)$ where span $(w_i, \ldots, w_j)$ refers to an entity mention. When using ELMo, we first encode the sentence: $(c_1, \ldots, c_T) = \text{ELMo}(w_1, \cdots, w_T)$, and we use the average of contextualized hidden states corresponding to the entity span as the contextualized entity representation. That is, $f_{\text{ELMo}}(w_{1:T}, i, j) = \frac{\sum_{k=i}^{j} c_k}{j-i+1}$.

With BERT, following Onoe and Durrett (2019), we concatenate the full sentence with the entity mention, starting with [CLS] and separating the two by [SEP], i.e., $[\text{CLS}], w_1, \ldots, w_T, [\text{SEP}], w_i, \ldots, w_j, [\text{SEP}]$. We encode the full sequence using BERT and use the output from the [CLS] token as the entity mention representation.

4.2 Encoders for Descriptive Entity Representations

We encode an entity description by treating the entity description as a sentence, and use the average of the hidden states from ELMo as the entity description representation. With BERT, we use the output from the [CLS] token as the description representation.

4.3 Hyperlink-Based Training

An entity mentioned in a Wikipedia article is often linked to its Wikipedia page, which provides a useful description of the mentioned entity. The same Wikipedia page may correspond to many different entity mentions. Likewise, the same entity mention may refer to different Wikipedia pages depending on its context. For instance, as shown in Figure 5, based on the context, "France" is linked to the Wikipedia page of "France national football team" instead of the country. The specific entity in the knowledge base can be inferred from the context information. In such cases, we believe Wikipedia provides valuable complementary information to the current pretrained CWRs such as BERT and ELMo.

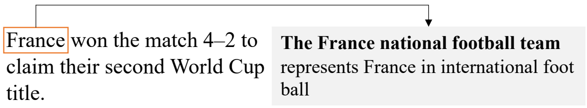

France won the match 4–2 to claim their second World Cup title.
The France national football team
represents France in international foot ball

Figure 5: An example of hyperlinks in Wikipedia. "France" is linked to the Wikipedia page of "France national football team" instead of the country France.

To incorporate such information during training, we automatically construct a hyperlink-enriched dataset from Wikipedia that we will refer to as WikiEnt. Prior work has used similar resources (Singh et al., 2012; Gupta et al., 2017), but we aim to standardize the process and will release the dataset.

The WikiEnt dataset consists of sentences with contextualized entity mentions and their corresponding descriptions obtained via hyperlinked Wikipedia pages. When processing descriptions, we only keep the first 100 word tokens at most as the description of a Wikipedia page; similar truncation has been done in prior work (Gupta et al., 2017). For context sentences, we remove those without hyperlinks from the training data and duplicate those with multiple hyperlinks. We also remove context sentences for which we cannot find matched Wikipedia descriptions. These processing steps result in a training set of approximately 92 million instances and over 3 million unique entities.

We define a hyperlink-based training objective and add it to ELMo. In particular, we use contextualized entity representations to decode the hyperlinked Wikipedia description, and also use the descriptive entity representations to decode the linked context. We use bag-of-words decoders in both decoding processes. More specifically, given a context sentence $x_{1:T_x}$ with mention span (i, j) and a description sentence $y_{1:T_y}$, we use the same bidirectional language modeling loss $l_{\text{lang}}(x_{1:T_x}) + l_{\text{lang}}(y_{1:T_y})$ in ELMo where

$$l_{\text{lang}}(u_{1:T}) = -\sum_{t=1}^{T} \log p(u_{t+1}|u_1, \ldots, u_t) + \log p(u_{t-1}|u_t, \ldots, u_T)$$

and p is defined by the ELMo parameters. In addition, we define the two bag-of-words reconstruction losses:

$$l_{\text{ctx}} = -\sum_t \log q(x_t | f_{\text{ELMo}}([\text{BOD}]y_{1:T_y}, 1, T_y))$$

$$l_{\text{desc}} = -\sum_t \log q(y_t | f_{\text{ELMo}}([\text{BOC}]x_{1:T_x}, i, j))$$

where [BOD] and [BOC] are special symbols prepended to sentences to distinguish descriptions from contexts. The distribution q is parameterized by a linear layer that transforms the conditioning embedding into weights over the vocabulary. The final training loss is

$$l_{\text{lang}}(x_{1:T_x}) + l_{\text{lang}}(y_{1:T_y}) + l_{\text{ctx}} + l_{\text{desc}} \qquad (1)$$

Same as the original ELMo, each log loss is approximated with negative sampling (Jean et al., 2015). We write EntELMo to denote the model trained by Eq. (1). When using EntELMo for contextualized entity representations and descriptive entity representations, we use it analogously to ELMo.

5 Experiments

5.1 Setup

As a baseline for hyperlink-based training, we train EntELMo on the WikiEnt dataset with only a bidirectional language model loss. Due to the limitation of computational resources, both variants of EntELMo are trained for one epoch (3 weeks time) with smaller dimensions than ELMo. We set the hidden dimension of each directional long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) layer to be 600, and project it to 300 dimensions. The resulting vectors from each layer are thus 600 dimensional. We use 1024 as the negative sampling size for each positive word token. For bag-of-words reconstruction, we randomly sample at most 50 word tokens as positive samples from the the target word tokens. Other hyperparameters are the same as ELMo. EntELMo is implemented based on the official ELMo implementation.[2]

As a baseline for contextualized and descriptive entity representations, we use GloVe word averaging of the entity mention as the "contextualized" entity representation, and use word averaging of the truncated entity description text as its description representation. We also experiment two variants of EntELMo, namely EntELMo w/o l_{ctx} and EntELMo with l_{etn}. For second variant, we replace l_{ctx} with l_{etn}, where we only decode entity mentions instead of the whole context from descriptions. We lowercased all training data as well as the evaluation benchmarks.

We evaluate the transferrability of ELMo, EntELMo, and BERT by using trainable mixing weights for each layer. For ELMo and EntELMo, we follow the recommendation from Peters et al. (2018a) to first pass mixing weights through a softmax layer and then multiply the weighted-summed representations by a scalar. For BERT, we find it better to just use unnormalized mixing weights. In addition, we investigate per-layer performance for both models in Section 6.

5.2 Results

Table 3 shows the performance of our models on the EntEval tasks. Our findings are detailed below:

- Pretrained CWRs (ELMo, BERT) perform the best on EntEval overall, indicating that they capture knowledge about entities in contextual mentions or as entity descriptions.
- BERT performs poorly on entity similarity and relatedness tasks. Since this task is zero-shot, it validates the recommended setting of finetuning BERT (Devlin et al., 2018) on downstream tasks, while the embedding of the [CLS] token does not necessarily capture the semantics of the entity.
- BERT Large is better than BERT Base on average, showing large improvements in ERT and NED. To perform well at ERT, a model must either glean particular relationships from pairs of lengthy entity descriptions or else leverage knowledge from pretraining about the entities considered. Relatedly, performance on NED is expected to increase with both the ability to extract knowledge from descriptions and by starting with increased knowledge from pretraining. The Large model appears to be handling these capabilities better than the Base model.
- EntELMo improves over the EntELMo baseline (trained without the hyperlinking loss) on some

[2] Our implementation is available at https://github.com/mingdachen/bilm-tf

	CAP	CERP	EFP	ET	ESR	ERT	NED	Average
GloVe	71.9	52.6	67.0	10.3	50.9	40.8	41.2	47.8
BERT Base	**80.6**	65.6	74.8	32.0	28.8	42.2	50.6	53.5
BERT Large	79.1	**66.9**	**76.7**	32.3	32.6	**48.8**	**54.3**	55.8
ELMo	80.2	61.2	75.8	**35.6**	60.3	46.8	51.6	**58.8**
EntELMo baseline	78.0	59.6	71.5	31.3	**61.6**	46.5	48.5	56.7
EntELMo	76.9	59.9	72.4	32.2	59.7	45.7	49.0	56.5
EntELMo w/o l_{ctx}	73.5	59.4	71.1	33.2	53.3	44.6	48.9	54.9
EntELMo w/ l_{etn}	76.2	60.4	70.9	33.6	49.0	42.9	49.3	54.6

Table 3: Performances of entity representations on EntEval tasks. Best performing model in each task is boldfaced. CAP: coreference arc prediction, CERP: contexualized entity relationship prediction, EFP: entity factuality prediction, ET: entity typing, ESR: entity similarity and relatedness, ERT: entity relationship typing, NED: named entity disambiguation. EntELMo baseline is trained on the same dataset as EntELMo but not using the hyperlink-based training. EntELMo w/ l_{etn} is trained with a modified version of l_{ctx}, where we only decode entity mentions instead of the whole context.

	Rare		CoNLL		ERT	
	Des.	Name	Des.	Name	Des.	Name
ELMo	38.1	36.7	63.4	71.2	46.8	31.5
BERT Base	42.2	36.6	64.7	74.3	42.2	34.3
BERT Large	48.8	44.0	64.6	74.8	48.8	32.6

Table 4: Accuracies (%) in comparing the use of description encoder (Des.) to entity name (Name).

	CoNLL
ELMo	71.2
Gupta et al. (2017)	65.1
Deep ED	66.7

Table 5: Accuracies (%) on CoNLL-YAGO with static or non-static entity representations.

tasks but suffers on others. The hyperlink-based training helps on CERP, EFP, ET, and NED. Since the hyperlink loss is closely-associated to the NED problem, it is unsurprising that NED performance is improved. Overall, we believe that hyperlink-based training benefits contextualized entity representations but does not benefit descriptive entity representations (see, for example, the drop of nearly 2 points on ESR, which is based solely on descriptive representations). This pattern may be due to the difficulty of using descriptive entity representations to reconstruct their appearing context.

6 Analysis

Is descriptive entity representation necessary? A natural question to ask is whether the entity description is needed, as for humans, the entity names carry sufficient amount of information for a lot of tasks. To answer this question, we experiment with encoding entity names by the descriptive entity encoder for ERT (entity relationship typing) and NED (named entity disambiguation) tasks. The results in Table 4 show that encoding the entity names by themselves already captures a great deal of knowledge regarding entities, especially for CoNLL-YAGO. However, in tasks like ERT, the entity descriptions are crucial as the names do not reveal enough information to categorize their relationships.

Table 5 reports the performance of different descriptive entity representations on the CoNLL-YAGO task. The three models all use ELMo as the context encoder. "ELMo" encodes the entity name with ELMo as descriptive encoder, while both Gupta et al. (2017) and Deep ED (Ganea and Hofmann, 2017) use their trained static entity embeddings. [3] As Gupta et al. (2017) and Deep ED have different embedding sizes from ELMo, we add an extra linear layer after them to map to the same dimension. These two models are designed for entity linking, which gives them potential advantages. Even so, ELMo outperforms them both by a wide margin.

Per-Layer Analysis. We evaluate each ELMo and EntELMo layer, i.e., the character CNN layer and two bidirectional LSTM layers, as well as each BERT layer on the EntEval tasks. Figure 6 reveals that for ELMo models, the first and second LSTM layers capture most of the entity knowledge from context and descriptions. The BERT layers show more diversity. Lower layers perform better on ESR (entity similarity and relatedness), while

[3] We note that the numbers reported here are not strictly comparable to the ones in their original paper since we keep all the top 30 candidates from Crosswiki while prior work employs different pruning heuristics.

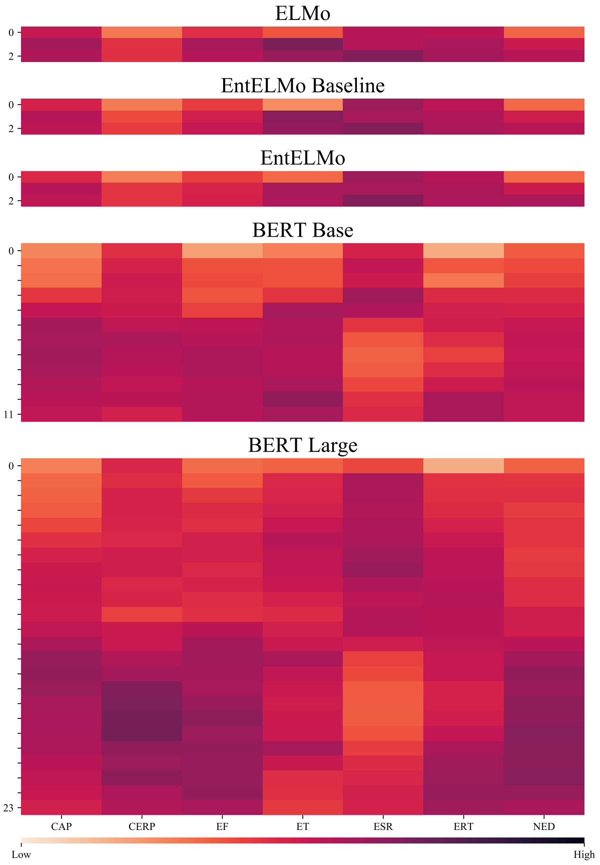


Figure 6: Heatmap showing per-layer performances for ELMo, EntELMo baseline, EntELMo, BERT Base, and BERT Large.

for other tasks higher layers are more effective.

7 Conclusion

Our proposed EntEval test suite provides a standardized evaluation method for entity representations. We demonstrate that EntEval tasks can benefit from the success of contextualized word representations such as ELMo and BERT. Augmenting encoding-decoding loss leveraging natural hyperlinks from Wikipedia further improves ELMo on some EntEval tasks. As shown by our experimental results, the contextualized entity encoder benefits more from this hyperlink-based training objective, suggesting future works to prioritize encoding entity description from its mention context.

Acknowledgments

This research was supported in part by a Bloomberg data science research grant to K. Stratos and K. Gimpel.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*.

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 534–545, Doha, Qatar. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019a. Evaluation benchmarks and learning criteriafor discourse-aware sentence representations. In *Proc. of EMNLP*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle:discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. FINET: Context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 868–878, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California. Association for Computational Linguistics.

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*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaochen Guo and Denilson Barbosa. 2014. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM ’14, pages 499–508, New York, NY, USA. ACM.

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

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–34, Sofia, Bulgaria. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hongzhao Huang, Larry Heck, and Heng Ji. 2015. Leveraging deep neural networks and knowledge graphs for entity disambiguation. *arXiv preprint arXiv:1504.07678*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. 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 (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2019. Boosting entity linking performance by leveraging unlabeled documents. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1935–1945, Florence, Italy. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. 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*, 3:315–328.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

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.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World knowledge for reading comprehension: Rare entity prediction with hierarchical lstms using external descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Shikhar Murty, Patrick Verga, Luke Vilnis, and Andrew McCallum. 2017. Finer grained entity typing with typenet. *arXiv preprint arXiv:1711.05795*.

Denis Newman-Griffis, Albert M. Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206, Melbourne, Australia. Association for Computational Linguistics.

Rasha Obeidat, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Description-based zero-shot fine-grained entity typing. 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 807–814, Minneapolis, Minnesota. Association for Computational Linguistics.

Yasumasa Onoe and Greg Durrett. 2019. Learning to denoise distantly-labeled data for entity typing. In *NAACL-HLT*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Maxim Rabinovich and Dan Klein. 2017. Fine-grained entity typing with high-multiplicity assignments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–334.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3168–3175, Istanbul, Turkey. European Language Resources Association (ELRA).

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2015. Corpus-level fine-grained entity typing using contextual information. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 715–725, Lisbon, Portugal. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *Transactions of the Association for Computational Linguistics*, 5(1):397–411.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.