EntEval: A Holistic Evaluation Benchmark for Entity Representations

Mingda Chen*[1], Zewei Chu*[2], Yang Chen[4], Karl Stratos[3], Kevin Gimpel[1]

*Equal Contribution. Listed in alphabetical order.

[1]Toyota Technological Institute at Chicago [2]University of Chicago [3]Rutgers University [4]Ohio State University

Learning Entity Representations

Entity $\xrightarrow{\text{Neural Networks}}$ Fixed-length vector

We are interested in two approaches:

- *Contextualized entity representations (CER)* that encode an entity based on the context it appears regardless of whether the entity is seen before.
- *Descriptive entity representations (DER)* that rely on entries in Wikipedia.

EntEval

- 7 probing task groups.

Entity Typing (ET)

ET = assign types to an entity given only the mention context.

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

Wisdom ✓ | University | Philosophy ✓ | Accident | …

Coreference Arc Prediction (CAP)

CAP = classify if two entities are the same given context

> Revenues of $14.5 billion were posted by [**Dell**]. [**The company**] ...
> ?

Entity Factuality Prediction (EFP)

EFP = classify the correctness of statements for entities.

> **TD Garden** has held Bruins games. ✓

Contexualized Entity Relationship Prediction (CP)

CP = classify the correctness of statements for entity pairs.

> **Gin and vermouth** can make **a martini**. ✓

EntEval cont.

Named Entity Disambiguation (NED)

NED = link a named-entity mention to its entry in a knowledge base.

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

A. China: China is a country in East Asia …

B. Porcelain: Porcelain is a ceramic material …

C. China_men's_national_basketball_team: The Chinese men's national basketball team represents the …

✓ D. China_PR_national_football_team: The Chinese national football team recognized as China PR by FIFA …

Entity Similarity and Relatedness (ESR)

ESR = predict the similarity of two entities given descriptions.

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

Entity Relationship Typing (ERT)

ERT = classify the types of relations between a pair of entities given descriptions.

book.school_or_movement.associated_works	
English Renaissance	Volpone

Statistics of EntEval

Task	Dataset	#class
NED	Rare	4
	CONLL-YAGO	≤ 30

Task	CAP	CP	EFP	ET	ESR	ERT
#class	2	2	2	10331	N/A	626

Dataset References

- ET: Ultra-fine entity typing.
- CAP: PreCo: A large-scale dataset in preschool vocabulary for coref resolution.
- CP: Conceptnet 5.5: An open multilingual graph of general knowledge.
- NED: Robust disambiguation of named entities in text.
- NED: Rare entity prediction with hierarchical lstms using external descriptions.
- ESR: Kore: keyphrase overlap relatedness for entity disambiguation.
- ESR: Jointly embedding entities and text with distant supervision.
- ERT: Freebase: a collaboratively created graph database for structuring human knowledge.

Hyperlink-Based Training

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 $l_{\text{lang}}(x_{1:T_x}) + l_{\text{lang}}(y_{1:T_y})$ as in ELMo, where

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

In addition, we define two bag-of-words reconstruction losses

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

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

Special symbols prepended to sentences to distinguish descriptions from contexts.

The final training loss for **EntELMo** is

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

Experiment Results

	ET	CAP	EFP	NED	CP	ERT	ESR
GloVe	10.3	71.9	67.0	41.2	52.6	40.8	50.9
BERT Base	32.0	**80.6**	74.8	50.6	65.6	42.2	28.8
BERT Large	32.3	79.1	**76.7**	**54.3**	**66.9**	**48.8**	32.6
ELMo	**35.6**	79.1	75.8	51.6	61.2	46.8	**60.3**

Table 1. Performances of entity representations on EntEval tasks.

	ET	CAP	EFP	NED	CP	ERT	ESR
EntELMo Baseline	31.3	**78.0**	71.5	48.5	59.6	**46.5**	**61.6**
EntELMo	32.2	76.9	**72.4**	49.0	59.9	45.7	59.7
EntELMo w/o l_{ctx}	33.2	73.5	71.1	48.9	59.4	44.6	53.3
EntELMo w/ l_{etn}	**33.6**	76.2	70.9	**49.3**	**60.4**	42.9	49.0

Table 2. EntELMo w/ l_{etn} is trained with a modified version of l_{ctx} where we only decode entity mentions instead of the whole context.

Static vs non-static entity representations

	CONLL-YAGO
ELMo	71.2
Gupta et al. 2017	65.1
Ganea and Hofmann, 2017	66.7

Scan to check out the code and data