Fine-Grained Entity Typing for Domain Independent Entity Linking

Yasumasa Onoe and Greg Durrett, {yasumasa, gdurrett}@cs.utexas.edu

1. Motivation

Here is an example from the CoNLL-YAGO dataset.

The Irish took a 4-0 lead within 20 minutes.

What does “Irish” mean here?
Irish people or Ireland national football team?

Entity linking training data contains mostly this…
…so linking model will never predict this

3. Entity Typing for Entity Linking

Mention & Context

If an attacker can trick someone inside into opening a Trojan horse, the malicious software can exploit the liberal egress policy...

Entity Typing Model

MENTION ENCODER
CONTEX ENCODER
ELMo/BERT +Bi-LSTM +Atten.
ELMo/BERT +Bi-LSTM +Atten.
Concat.
LINEAR+SIGMOID

Typing model encodes mention and context and predicts types independently using binary classification from a shared representation.

2. Our Approach

Our model predicts fine-grained entity types of “Irish” that make sense in this context. Supervision comes from mention-types, and this way prevents a model from memorizing mention-entity pairs. As a result, our model generalizes well to new domains.

4. Training

Entity Typing Training Data

For each hyperlink in Wikipedia, treat the text (New York City) as a mention and the destination page’s Wiki categories (Cities, Populated places, …) as gold entity types.

Type Set

We derive a type set from Wikipedia categories. To obtain more coarse-grained types, we split each category using several rules. We use a vocabulary of 60,000 types in our experiments.

5. Experiments

CoNLL-YAGO (standard benchmark dataset)

- Significant entity overlap between standard training and standard test set

Baselines:
- Popularity prior baseline (Most Frequent);
- Neural baseline trained on Wikipedia (CBoW+w2v);
- SOTA models that do not use in-domain training data (Gupta+ 17, Lazic+ 15).

Results: Without using the in-domain training data, our approach shows strong performance on this dataset. Supervised systems (Raiman+) still achieve stronger performance, but much of this is due to entity memorization.

WikilinksNED: Unseen-Mentions (challenging setting)

- All mentions in development set do not occur in test set

Baselines:
- Popularity prior baseline (Most Frequent);
- SOTA neural EL model (Mueller+ 18) trained on either WikilinksNED or Wikipedia data.

Results: Our approach handles unseen mention-entity pairs better. State-of-the-art models do poorly when generalization to new entities is required.

Code & data available at https://github.com/yasumasaonoe/ET4EL