Neural Fine-Grained Entity Type Classification with Hierarchy-Aware Loss

Peng Xu

Denilson Barbosa

Department of Computing Science
University of Alberta

The Task: Fine-Grained Entity Type Classification

- Traditional *Coarse-Grained Entity Type Classification*, as a sub-task of *Named Entity Recognition (NER)*, focuses on a small set of coarse types.
- *Fine-Grained Entity Type Classification (FETC)* aims at labeling entity mentions in context with one or more specific types organized in a hierarchy.

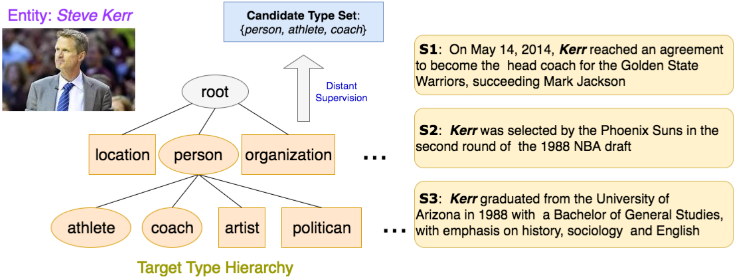
Figure: Traditional coarse-grained types are colored in black. Fine-grained types are colored in red.

Motivation

Fine-grained types help in many applications:

- relation extraction
- question answering
- coreference resolution
- entity linking
- knowledge base completion
- entity recommendation
- and so on...

Characteristics of FETC



- Context dependent labeling
- Hierarchical structure of entity types
- Collapse of the mutual exclusion assumption
- Noise in automatically annotated data

Multi-label vs. Single label

In FETC, types are not mutually exclusive!
It is natural to formulate the task as a multi-label classification problem and most FETC methods adopt this setting.
However,

- context dependent labeling $\rightarrow$ assumption that one mention can only have one *type-path* along the hierarchy
- type hierarchy is a tree $\rightarrow$ each *type-path* can be uniquely represented by the terminal type (not necessarily a leaf node)

Then, the task can be transformed to predict the terminal type of the *type-path* in the hierarchy, which is a single-label classification problem!

Pros and Cons to Adopt Single Label Setting

Pros:

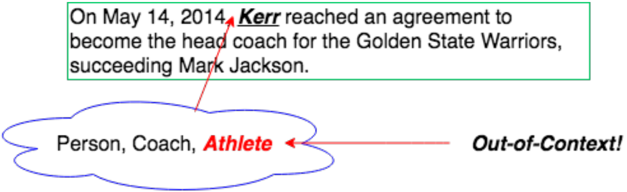
1. Simpler and more elegant
2. Benefit from previous research
3. No post-processing needed

Cons:

1. The upper bounds are no longer 100% (But, is that really important? State-of-the-art methods are nowhere near 80% strict accuracy.)

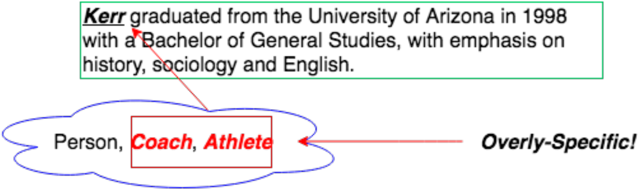
	FIGER(GOLD)	OntoNotes
# types	113	89
# raw testing mentions	563	8963
% testing mentions with single *type-path*	88.28	94.00

Out-of-context Noise



One kind of noise introduced by distant supervision is assigning labels that are *out-of-context*.

Overly-specific Noise



Another source of noise introduced by distant supervision is when the type is *overly-specifc* for the context.

Typical FETC Methods

	Attentive	AFET	LNR	A&A
without manual features	✗	✗	✗	✓
use attentive neural network	✓	✗	✗	✗
adopt single label setting	✗	✗	✗	✗
handle *out-of-context* noise	✗	✓	✓	✓
handle *overly-specifc* noise	✗	✓	✓	✗

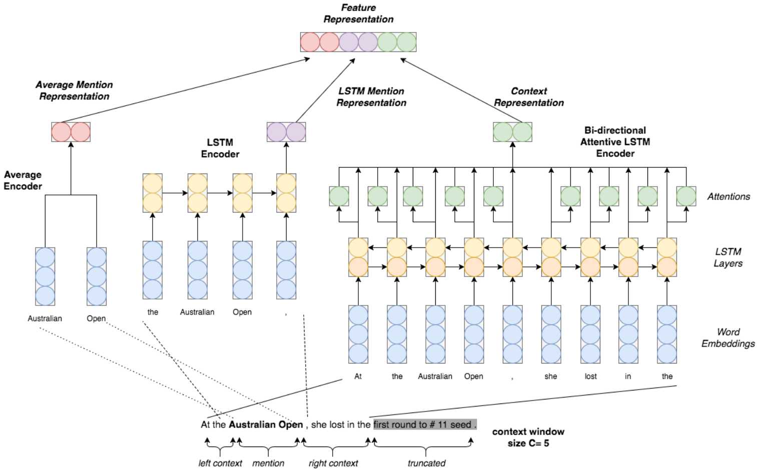
1. Attentive: Shimaoka *et al.* (2017)
2. AFET: Ren *et al.* (2016a)
3. LNR: Ren *et al.* (2016b)
4. A&A: Abhishek and Awekar (2017)

Our Proposed Model: NFETC

NFETC is a single, much simpler and more elegant neural model that attempts FETC "end-to-end" without post-processing or ad-hoc features.

	Attentive	AFET	LNR	A&A	NFETC
without manual features	✗	✗	✗	✓	✓
use attentive neural network	✓	✗	✗	✗	✓
adopt single label setting	✗	✗	✗	✗	✓
handle *out-of-context* noise	✗	✓	✓	✓	✓
handle *overly-specifc* noise	✗	✓	✓	✗	✓

Neural Architecture



A Simple Yet Effective Variant of Cross-Entropy

Traditional cross-entropy loss:

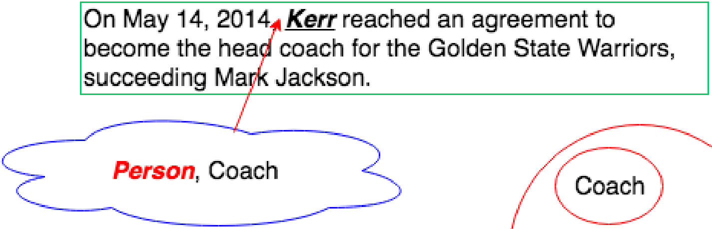
$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(\hat{p}(y_i)), \tag{1}$$

which can't handle data with multi *type-paths* (that is, with *out-of-context* noise). A simple yet effective variant of the cross-entropy loss:

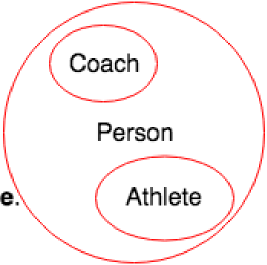
$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(\hat{p}(y_i^*)), \tag{2}$$

where $y_i^* = \arg\max_{y \in \mathcal{Y}_i} \hat{p}(y)$ and $\mathcal{Y}_i$ is the labelled type set.

Hierarchical Loss Normalization: Intuition



What if we predict ***Kerr*** as **Person** here?
It's correct in some sense compared to **Athlete**.
Types are correlated!



Hierarchical Loss Normalization [1]

Based on the intuition, we adjust the estimated probability to:

$$p^*(\hat{y}) = p(\hat{y}) + \beta * \sum_{t \in \Gamma} p(t) \tag{3}$$

where Γ is the set of ancestor types along the *type-path* of $\hat{y}$, β is a hyperparameter. In this way, the model will:

1. get less penalty when it predicts the actual type for data with *overly-specific* noise
2. prefer generic types unless there is a strong indicator for a more specific type in the context

[1] Hierarchical loss function (Cai and Hofmann, 2004) was originally introduced in the context of document categorization with SVM. However, they assume that weights to control the hierarchical loss can be solicited from domain experts which is inapplicable for FETC.

Experiments

- Datasets: FIGER(GOLD) & OntoNotes

	FIGER(GOLD)	OntoNotes
# types	113	89
# raw training mentions	2009898	253241
# raw testing mentions	563	8963
% filtered training mentions	64.46	73.13
% filtered testing mentions	88.28	94.00
Max hierarchy depth	2	3

- Evaluation Metrics: Strict Accuracy, Macro F1 and Micro F1

Results

Model	FIGER(GOLD) Strict Acc.	Macro F1	Micro F1	OntoNotes Strict Acc.	Macro F1	Micro F1
Attentive	59.68	78.97	75.36	51.74	70.98	64.91
AFET	53.3	69.3	66.4	55.1	71.1	64.7
LNR+FIGER	59.9	76.3	74.9	57.2	71.5	66.1
A&A	65.8	**81.2**	77.4	52.2	68.5	63.3
NFETC(f)	57.9 ± 1.3	78.4 ± 0.8	75.0 ± 0.7	54.4 ± 0.3	71.5 ± 0.4	64.9 ± 0.3
NFETC-hier(f)	68.0 ± 0.8	**81.4 ± 0.8**	77.9 ± 0.7	59.6 ± 0.2	**76.1 ± 0.2**	69.7 ± 0.2
NFETC(r)	56.2 ± 1.0	77.2 ± 0.9	74.3 ± 1.1	54.8 ± 0.4	71.8 ± 0.4	65.0 ± 0.4
NFETC-hier(r)	**68.9 ± 0.6**	**81.9 ± 0.7**	**79.0 ± 0.7**	**60.2 ± 0.2**	**76.4 ± 0.1**	**70.2 ± 0.2**

Variants of our proposed model:

- NFETC(f): basic model trained on data with single *type-path*
- NFETC-hier(f): with hierarchical loss normalization trained on data with single *type-path*
- NFETC(r): with variant of cross-entropy trained on raw data
- NFETC-hier(f): with variant of cross-entropy and hierarchical loss normalization trained on raw data

Case Study

Test Sentence	Ground Truth	Prediction (w/o HLN)	Prediction (w HLN)
S1: **Hopkins** said four fellow elections is curious , considering the . . .	**Person**	**Politician**	**Person**
S2: . . . for WiFi communications across all **the SD cards.**	**Product**	**Software**	**Product**

Test Sentence	Ground Truth	Prediction (original CE)	Prediction (improved CE)
S3: **ASC** Director Melvin Taing said that because the commission is . . .	**Organization**	**Title**	**Organization**

Test Sentence	Ground Truth	Failed Prediction
S4: A handful of professors in the **UW** Department of Chemistry . . .	**Educational Institution**	**Organization**
S5: Work needs to be done and, in **Washington state**, . . .	**Province**	**City**

Conclusion

- Studied two kinds of noise, namely *out-of-context* noise and *overly-specific* noise.
- Propose a neural model which jointly learns representations for entity mentions and their context.
- A variant of cross-entropy loss function was used to handle *out-of-context* noise.
- Hierarchical loss normalization was introduced to alleviate the negative effect of *overly-specific* noise.
- Outperform previous state-of-the-art methods significantly.

Q&A

Questions?

Homepage

Code