Multi-Task Transfer Learning for Fine-Grained Named Entity Recognition

Masato Hagiwara[1], Ryuji Tamaki[2], Ikuya Yamada[2]

[1]Octanove Labs

[2]Studio Ousia

Named Entity Recognition (NER)

- Few systems deal with more than 100+ types
  - cf. FIGER 112 types (Ling and Weld, 2012)
- Entity typing
  - (Ren et al., 2016), (Shimaoka et al., 2016), (Yogatama et al., 2015)

Can we solve NER (detection and classification) with 7,000+ types in a generic fashion?

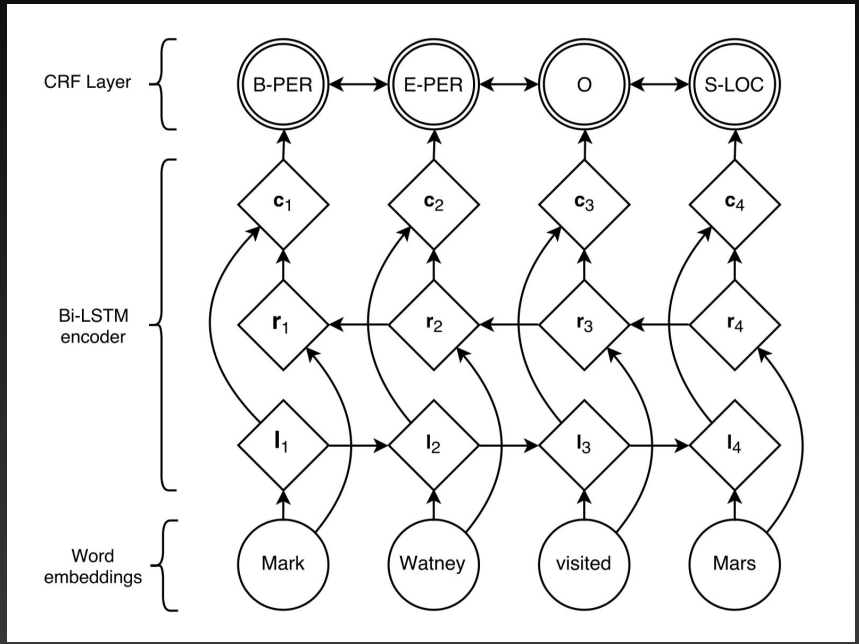
Challenge 1: Lack of Training Data

Lack of NER datasets annotated with AIDA

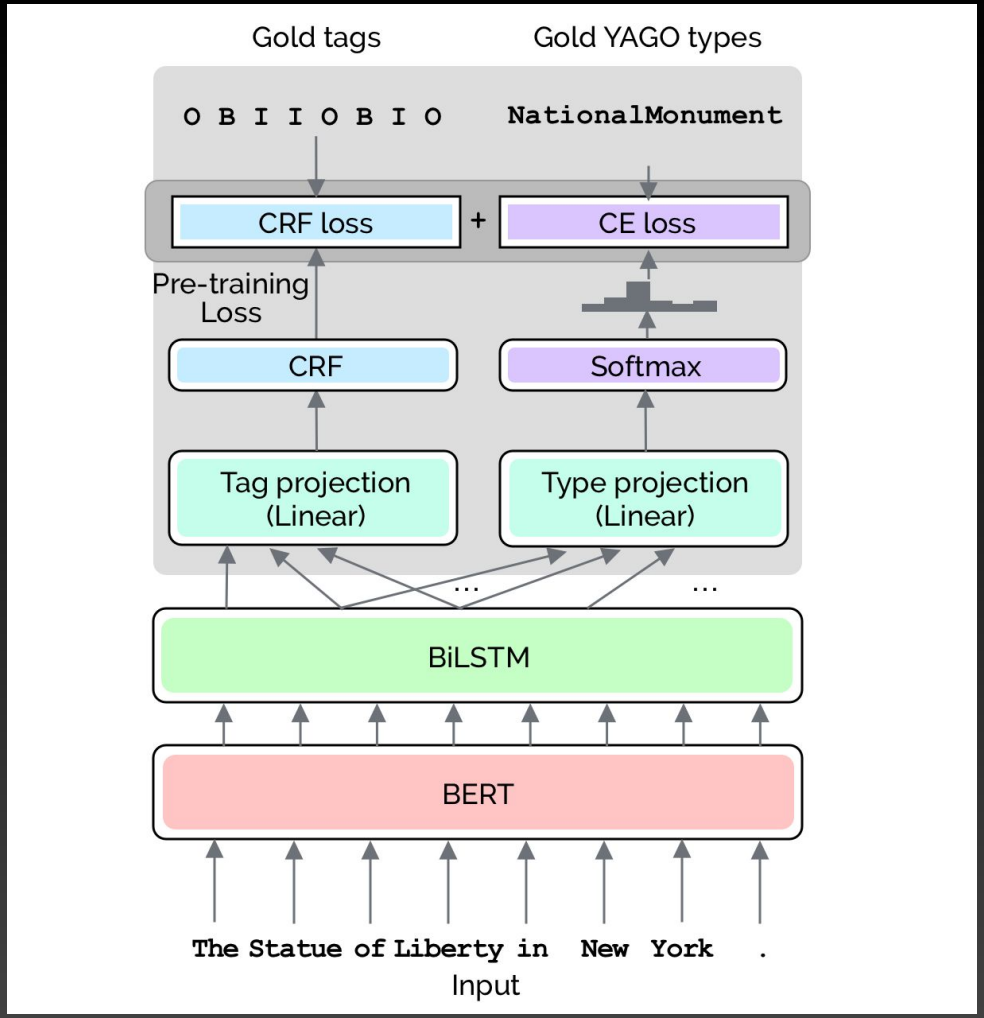
Silver-standard dataset with YAGO annotations

Transfer learning to AIDA

Challenge 2: Large Tag Set



Cost of CRF = $O(n^2)$ (n = # of types)


Gold tags
Gold YAGO types
O B I I O B I O
NationalMonument
CRF loss
+
CE loss
Pre-training Loss
CRF
Softmax
Tag projection (Linear)
Type projection (Linear)
...
...
BiLSTM
BERT
The Statue of Liberty in New York .
Input

Challenge 3: Ambiguity in Types

House103544360

vs

House107971449

WorldOrganization108294696

vs

Alliance108293982

Plaza108619795

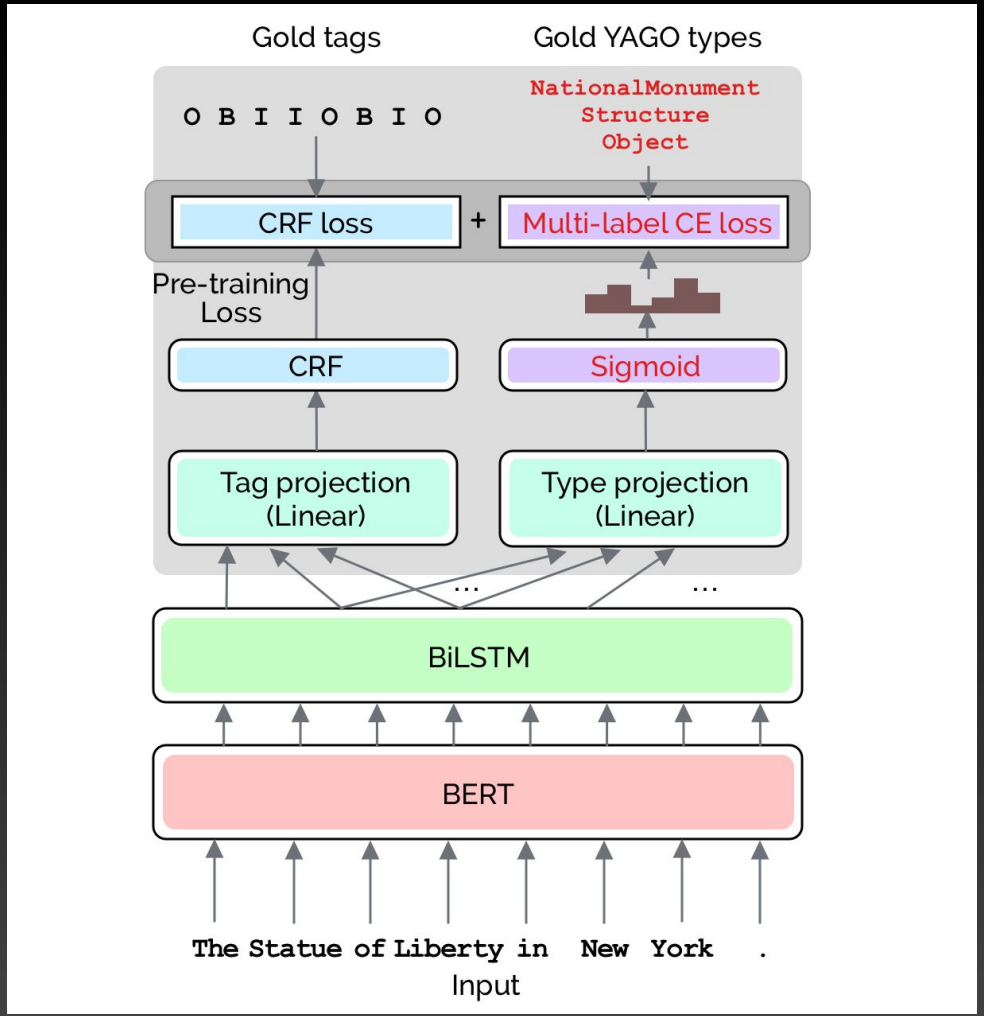
vs

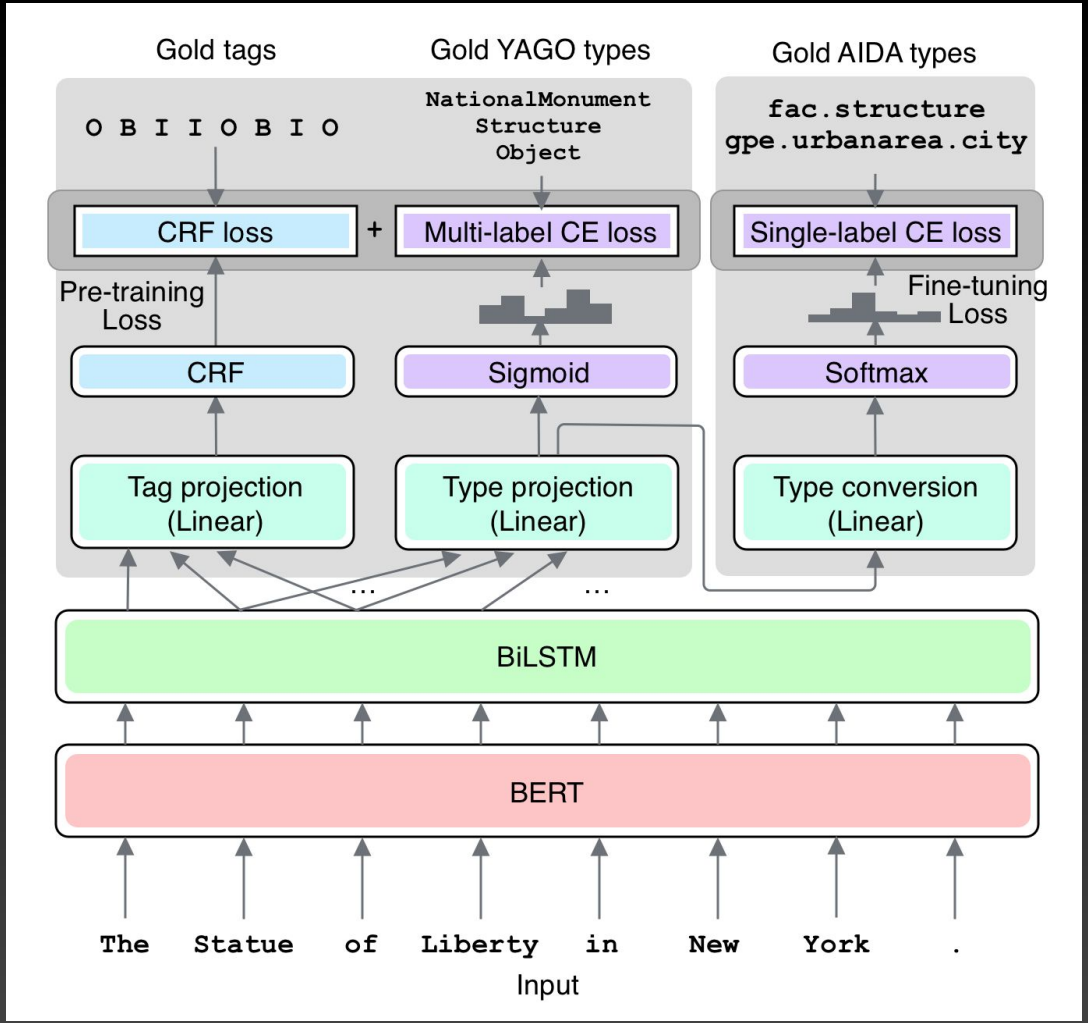
Plaza103965456

Hierarchical Multi-label Classification

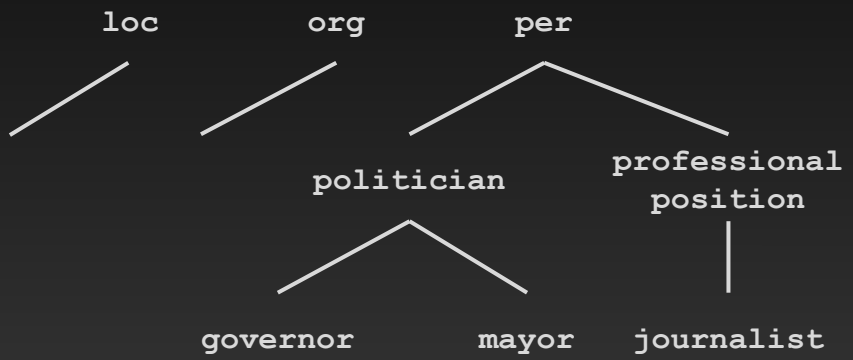
PhysicalEntity	YagoGeoEntity
Object	Location
Whole	Region
Artifact	District
Structure	AdministrativeDistrict
Memorial	Municipality
NationalMonument	City

The Statue of Liberty in New York


Gold tags
Gold YAGO types
O B I I O B I O
NationalMonument
Structure
Object
CRF loss
+
Multi-label CE loss
Pre-training
Loss
CRF
Sigmoid
Tag projection
(Linear)
Type projection
(Linear)
...
...
BiLSTM
BERT
The Statue of Liberty in New York .
Input

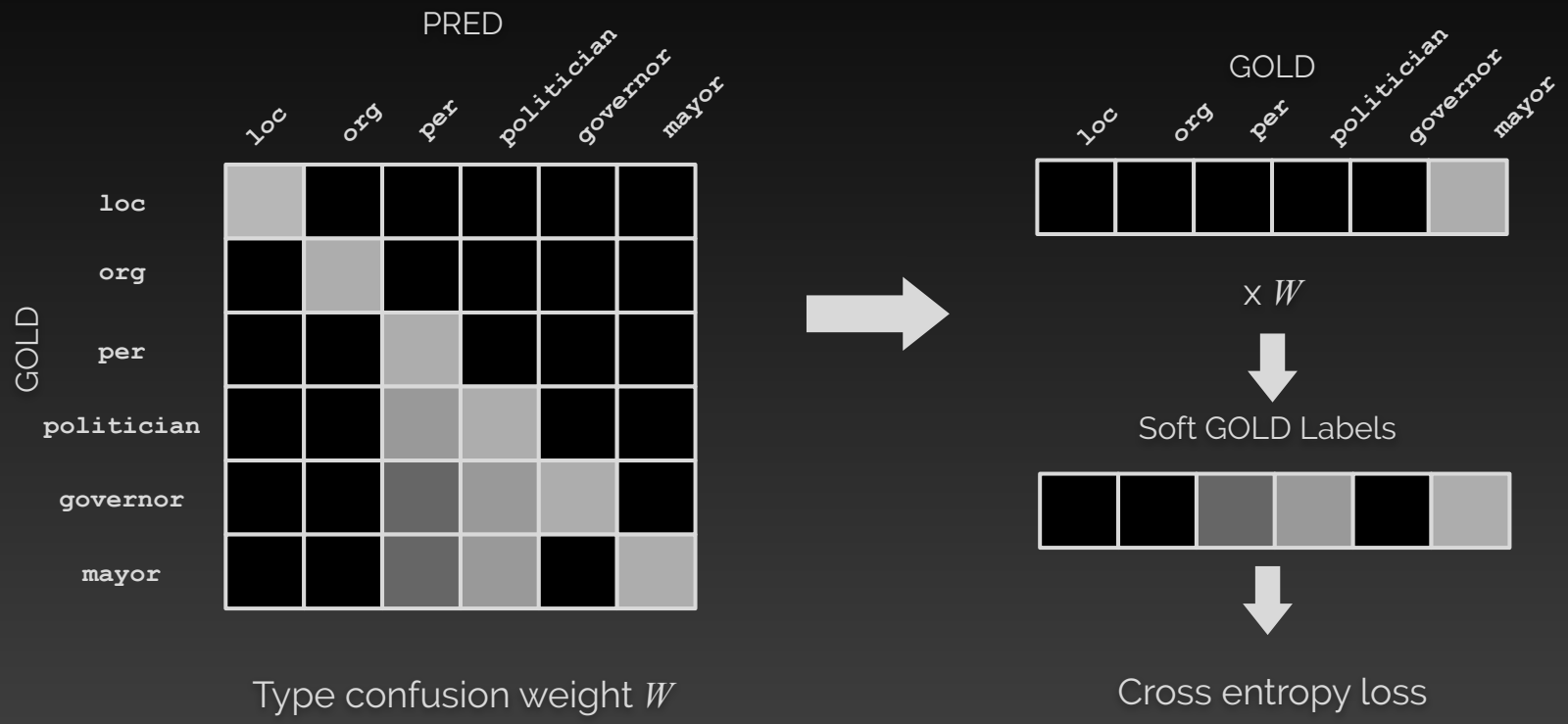

Gold tags
Gold YAGO types
Gold AIDA types
O B I I O B I O
NationalMonument
Structure
Object
fac.structure
gpe.urbanarea.city
CRF loss
+
Multi-label CE loss
Single-label CE loss
Pre-training
Loss
Fine-tuning
Loss
CRF
Sigmoid
Softmax
Tag projection
(Linear)
Type projection
(Linear)
Type conversion
(Linear)
...
...
BiLSTM
BERT
The
Statue
of
Liberty
in
New
York
.
Input

Challenge 4: Hierarchical Types



Hierarchy-aware soft loss

Hierarchy-Aware Soft Loss



Experiments

Datasets

1) Pre-training
   OntoNotes 5.0 (subset) for detection
   Silver-standard Wikipedia for classification
   Manually-annotated subset for dev.

2) Fine-tuning
   Manually-annotated WIkipedia
   Manually-fixed AIDA Source Data
   Manually-annotated OntoNotes 5.0 (subset)

Settings

- Embeddings
  bert-base-cased
  2-layer BiLSTM (200 hidden units)

- Type conversion
  2-layer feed-forward with ReLU

- Optimization
  Adam (lr = 0.001) for pre-training
  BertAdam (lr = 1e-5 with 2,500 warm-up)

Results

TEAM	F1
Maximum	0.614
OusiaNER (ours)	0.499
Median	0.423

Method	Prec	Rec	F1
Direct	0.45	0.42	0.43
Fine-tuned	**0.65**	**0.57**	**0.61**
Fine-tuned w/o loss	0.60	0.50	0.55

Error Analysis

- Location vs GPE
  - “Southern Maryland”
    OK: **`loc.position.region`**, NG: **`gpe.provincestate.provincestate`**
- Ethnic/national groups
  - “Syrians”
    OK: no annotation, NG: **`gpe.country.country`**
- Type too specific
  - “Obama”
    OK: **`per.politician`**, NG: **`per.politician.headofgovernment`**
- Type too generic
  - “SANA news agency”
    OK: **`org.commercialorganization.newsagency`**, NG: **`org`**

Conclusion

- Multi-task transfer learning approach for ultra fine-grained NER
  - Transfer learning from YAGO to AIDA
  - Multi-task learning of named entity detection and classification
  - Multi-label classification of named entity types
  - Hierarchy-aware soft loss

Improvement Ideas

- Using “type name” embeddings
  - e.g., **`per.professionalposition.spokesperson`**
  - e.g., **`org.commercialorganization.newsagency`**
- Gazetteers and handcrafted features
- Hierarchical model
  - BIO+loc/org/per/... -> more fine-grained types
- Ensemble
- Post-processing
- Finally... read the annotation guideline and examine the training data!

Thanks for listening!



Masato Hagiwara[1], Ryuji Tamaki[2], Ikuya Yamada[2]

[1]Octanove Labs
http://www.octanove.com/

[2]Studio Ousia
http://www.ousia.jp/en/