TRANSFORMER-BASED METHODS FOR RECOGNIZING ULTRA FINE-GRAINED ENTITIES (RUFES)

Authors: **Emanuela Boros** and Antoine Doucet

TAC 2020 Workshop

February 22-23, 2021

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

Contents

1. **Introduction**
2. **Dataset**
3. **Ultra Fine-grained Entities**
   - 3.1 Data Pre-processing
   - 3.2 **Entity Extraction** Model
   - 3.2 Within-document **Entity Coreference** Resolution Model
   - 3.3 Experiments & Results
4. **Conclusions & Perspectives**

Introduction

- Fine-grained entity recognition: labeling entity mentions in context with one or more specific types organized in a hierarchy

  $Photographer \in Artist \in Person$

- Two phases:
  - a preliminary phase where the data is provided along with a limited annotated set of samples (50 documents)
  - human feedback was provided for the preliminary submissions based on a user model of how analysts might interact with the systems

Dataset

Dataset

KBP 2020 RUFES dataset

- follows the three-level x.y.z hierarchy
- 200 fine-grained entity types

  → course-level entity types (14), **APP**, **FAC**, **LOC**, etc.

  → fine-grained entity types, Publication.**Magazine.NewsMagazine**, APP.**CommunicationSoftware.SocialMedia**, etc.
- 100,000 development source documents
- 50 annotated documents
- 100,000 the evaluation source documents

Ultra Fine-grained Entities Methods

We separated RUFES in two sub-tasks:

- **Entity extraction**: the detection and the classification of fine-grained entity types including the named, nominal, and pronominal mentions for each mention (labeled as NAM, NOM, and PRO, respectively);
- **Within-document entity coreference resolution**: the detection of the referential mentions in a document that point to the same entity.

The provided data was organized into two formats:

- *./rsd/*: “raw source data” (rsd) plain text form of the new article
- *./ltf/*: “logical text format” (ltf) derived from the rsd version, fully segmented and tokenized version of the corresponding rsd

TOKEN	FIRST_LEVEL	SECOND_LEVEL	THIRD_LEVEL	FORTH_LEVEL
Georgetown	B-ORG	B-EducationalInstitution	B-College	B-NAM
University	I-ORG	I-EducationalInstitution	I-College	I-NAM
officials	B-PER	B-Professional	O	B-NOM
on	O	O	O	O
Thursday	O	O	O	O
announced	O	O	O	O
that	O	O	O	O
they	B-ORG	B-EducationalInstitution	B-College	B-NOM
would	O	O	O	O
build	O	O	O	O
additional	O	O	O	O
on	O	O	O	O
-	O	O	O	O
campus	O	O	O	O
student	O	O	O	O
housing	B-FAC	B-Building	B-ApartmentBuilding	B-NOM

Figure 1: Data formatting example for the KBP 2020 RUFES dataset.

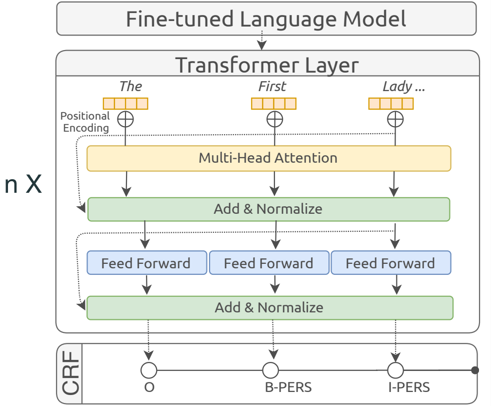


Figure 2: BERT-based model and the additional Transformer layers proposed by *Boros, Hamdi et al., 2020; Boros, Pontes et al., 2020*.

- Pre-trained and Fine-tuned language model
- BERT is a bidirectional stack of Transformer encoders
    - Masked Language Model
    - Next Sentence Prediction
- n×Transformer:
    - stack of identical layers: multi-head self-attention mechanism + position-wise fully connected feed-forward network
- multitask (coarse + fine)
- *bert-large-cased* + 2 × Transformer + CRF

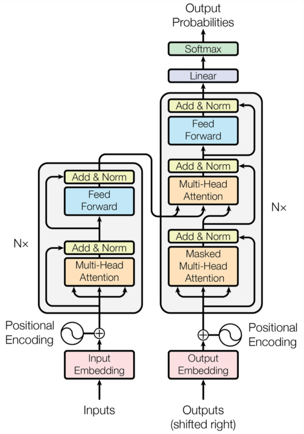


Figure 3: Transformer architecture [Vaswani et al., 2017].

Ultra Fine-grained Entity Extraction Model

multitask learning $\leftarrow$ this method has a label independence assumption $\leftarrow$ not valid for fine-grained entity extraction

$\rightarrow$ following the three-level x.y.z hierarchy, offering more confidence to the last predicted entity subtype (.z)

GPE.ProvinceState $\rightarrow$ check the ontology $\rightarrow$ **ProvinceState** $\notin$ **GPE** $\rightarrow$ LOC.**ProvinceState**

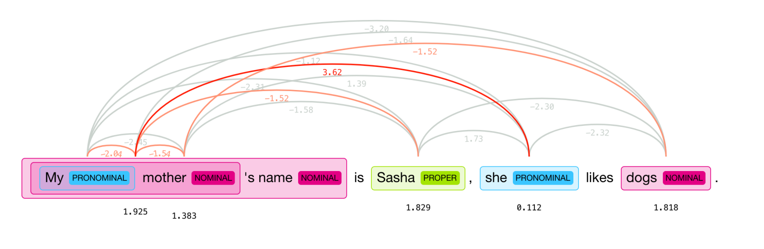
ORG.CommercialOrganization.**SocialMedia** $\rightarrow$ check the ontology $\rightarrow$ **SocialMedia** $\notin$ **CommercialOrganization** $\rightarrow$ APP.CommunicationSoftware.**SocialMedia**

NeuralCoref `https://github.com/huggingface/neuralcoref`

- previously trained on OntoNotes 5.0 dataset

  `https://www.gabormelli.com/RKB/OntoNotes_Corpus`

  - a rule-based mentions detection module (spaCy) to identify a set of potential coreference mentions;
  - a feed-forward neural-network which computes a coreference score for each pair of potential mentions



- applied in a within-document context

Experiment & Results

Preliminary Phase

- 1-first-rufes submission *bert-large-cased* + 2 × Transformer + CRF without coreference
- 2-first-rufes submission *bert-large-cased* + 2 × Transformer + CRF

After Feedback Phase

- 1-feedback-rufes & 2-feedback-rufes submissions = 2-first-rufes + Rule-based Feedback Inclusion

Feedback Inclusion

→ the first 40 errors detected in 10 random documents were reported

- 46% **wrong type** (mention-level entity types that do not exactly match the gold mention-level entity types)
- 12% **missing mentions**
- 11% **extraneous mentions** (a mention span does not exactly match or overlap with any gold mention span)
- 11% **wrong entity coreference**, either missing, incorrect or spurious
- 5% **wrong extents** (a mention span and gold mention span overlap but have different extents)

Feedback Inclusion

$\rightarrow$ wrong type errors

- related to entities that had one of the ontology terms included in the entity
- *"Norovirus"* was recognized as GPE (geopolitical entity) instead of Pathogen.Virus, *"virus"* $\in$ *"Norovirus"* & *"virus"* $\in$ **Pathogen.Virus**
- $\forall$ entities that included a fine-grained ontology type (level .z from x.y.z) i.e. *"Airport"*, *"Hospital"*, *"Highway"*, a rule was created to change the predictions into the correct types

The scoring results (F-score) for RUFES 2020 evaluation for all our submissions

Submission	strong mention match	strong typed mention match	mention ceaf	typed mention ceaf	entity ceaf	fine grain typing
1-first-rufes	0.868	0.745	0.552	0.503	0.551	0.3188
2-first-rufes	0.868	0.745	0.578	0.503	0.567	0.3188
1-feedback-rufes	0.868	0.745	0.578	0.504	0.567	0.3204
2-feedback-rufes	0.868	0.745	0.578	0.504	0.567	0.3239
Median	0.805	–	–	–	0.578	0.2313
Maximum	0.868	–	–	–	0.689	0.4162

Table 1: Median and Maximum scores are computed on the best-performing submission from each participant, as shared by RUFES organizers.

Conclusions

Conclusions

- The **BERT+$n\times$Transformer** has great potential for identifying ultra fine-grained entity types
- **BERT-alone** in comparison with BERT+**n**$\times$**Transformer** creates **more spurious cases** *Boros, Hamdi et al., 2020*
- $n\times$**Transformer** > 2 could lead to **overfitting**
- This type of model appears to be adapted for fine-grained entity extraction but we propose to refine the model in order to be able to take into consideration the **inter-dependencies between entity types**
- Improving the **entity coreference model** (re-training, etc.)
- **Further analysis remains to be done**

Thank you for your attention!

Emanuela Boros: emanuela.boros@univ-lr.fr

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).