Studio Ousia at the NTCIR-15 SHINRA2020-ML Task

Sosuke Nishikawa, Ikuya Yamada

Studio Ousia

Abstract

We solve the task as multi-class text classification based on **text-based feature** and **entity-based features** extracted from Wikipedia descriptions.

Materials & Methods

Extract text-based feature and entity-based features from an entity and its description obtained from Wikipedia.

Text-based feature

Feed entity descriptions into XLM-RoBERTa [1]
→ Use the output embedding corresponding to the [CLS] input token.

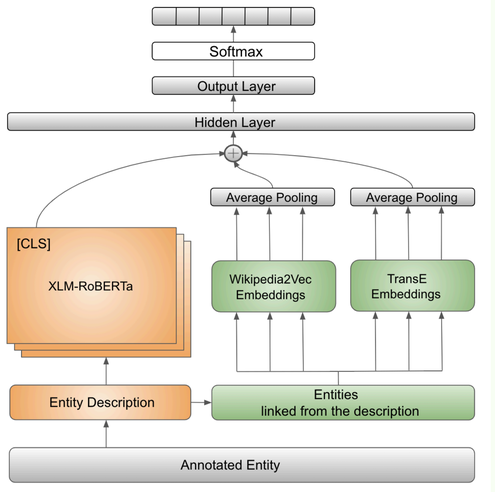
Entity-based features

Convert entities to following embeddings:
1: Wikipedia2Vec [2]
2: TransE model embedding (PyTorch-BigGraph [3])

→ Use element-wise average of these embeddings.

→ Concatenate these features and pass them to a hidden layer and an output layer with softmax function.

Fig. 1: Our model architecture



Heuristic Approach

Several entity pairs frequently co-occur
→ If our model predicts an entity type contained in one of the extracted pairs, we add the other type to the prediction.

Table 1: The top 10 frequent label pairs

label pairs		num
Ship	Weapon	6503
Archaeological_Place_Othe	Castle	1428
Company	Channel	1200
Line_Other	Car	1123
Shopping_Complex	Car_Stop	1080
Aircraft	Weapon	1034
Vehicle_Other	Weapon	586
Water_Route	Ship	410
Organization_Other	Channel	399
Company	Product_Other	353

Data augmentation

Use annotated Japanese Wikipedia data as extra training data.

Results

Preliminary experiment
- Use XLM-RoBERTa base
- Only in German

Final submissions
- Use XLM-RoBERTa large

Table 2: Results of preliminary experiments and final submissions

Model name	Precision	Recall	Micro F1
XLM-RoBERTa$_{base}$	0.713	0.713	0.713
XLM-RoBERTa$_{base}$ + Wikipedia2Vec	0.724	0.724	0.724
XLM-RoBERTa$_{base}$ + Wikipedia2Vec + TransE	0.725	0.725	0.725
XLM-RoBERTa$_{base}$ + Wikipedia2Vec + TransE (+Japanese)	0.739	0.741	0.739

Language	Micro F1	Rank
Arabic	70.52	3
German	**81.86**	**1**
Spanish, Castilian	80.94	2
French	**81.01**	**1**
Hindi	69.75	3
Italian	81.21	4
Portuguese	81.40	3
Thai	76.36	3
Chinese	79.76	2

- Incorporating all of features results in improved performance.
- achieves the highest scores of 81.86 in German and 81.01 in French.

[1] Conneau et al. Unsupervised Cross-lingual Representation Learning at Scale In ACL, 2020
[2] https://wikipedia2vec.github.io/wikipedia2vec/
[3] https://github.com/facebookresearch/PyTorch-BigGraph