

Priberam Labs at the NTCIR-15 SHINRA2020-ML: Classification Task



Rúben Cardoso, Afonso Mendes, Andre Lamurias
 Priberam Labs, Portugal
 ruben.cardoso@priberam.com

1. SHINRA2020-ML: Wikipedia entities classification

- Categorise Wikipedia entities based on the Extended Named Entity taxonomy
- Problem of multilingual multi-label classification
- We propose 3 models based on Multilingual BERT's (mBERT) embeddings
- Only the first 511 tokens of each Wikipedia page are leveraged

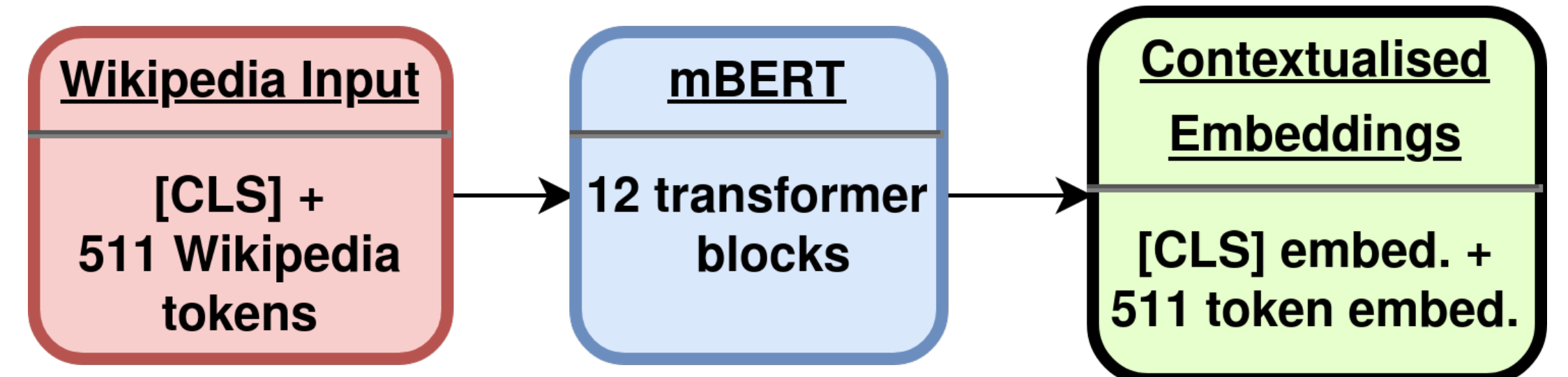


Figure 1: Multilingual BERT provides the contextualised embeddings.

2. Training Data

- 3.1M annotated Wikipedia pages across 13 languages
- English (EN), German (DE), Spanish (ES), French (FR), Italian (IT), Portuguese (PT), Russian (RU), Turkish (TR), Arabic (AR), Chinese (ZH), Polish (PL), Dutch (NL), and Korean (KO)

3. Linear Classification

- Linear layer projects mBERT's pooled representation onto the decision space
- Hierarchical structure not explicitly leveraged, only leaf labels are considered
- 3 pooling strategies were tested:
 - Linear+CLS: only leverages [CLS] embedding
 - Linear+Mean: mean of [CLS] + 511 tokens embeddings
 - Linear+Concat: combines mean and concatenation of embeddings

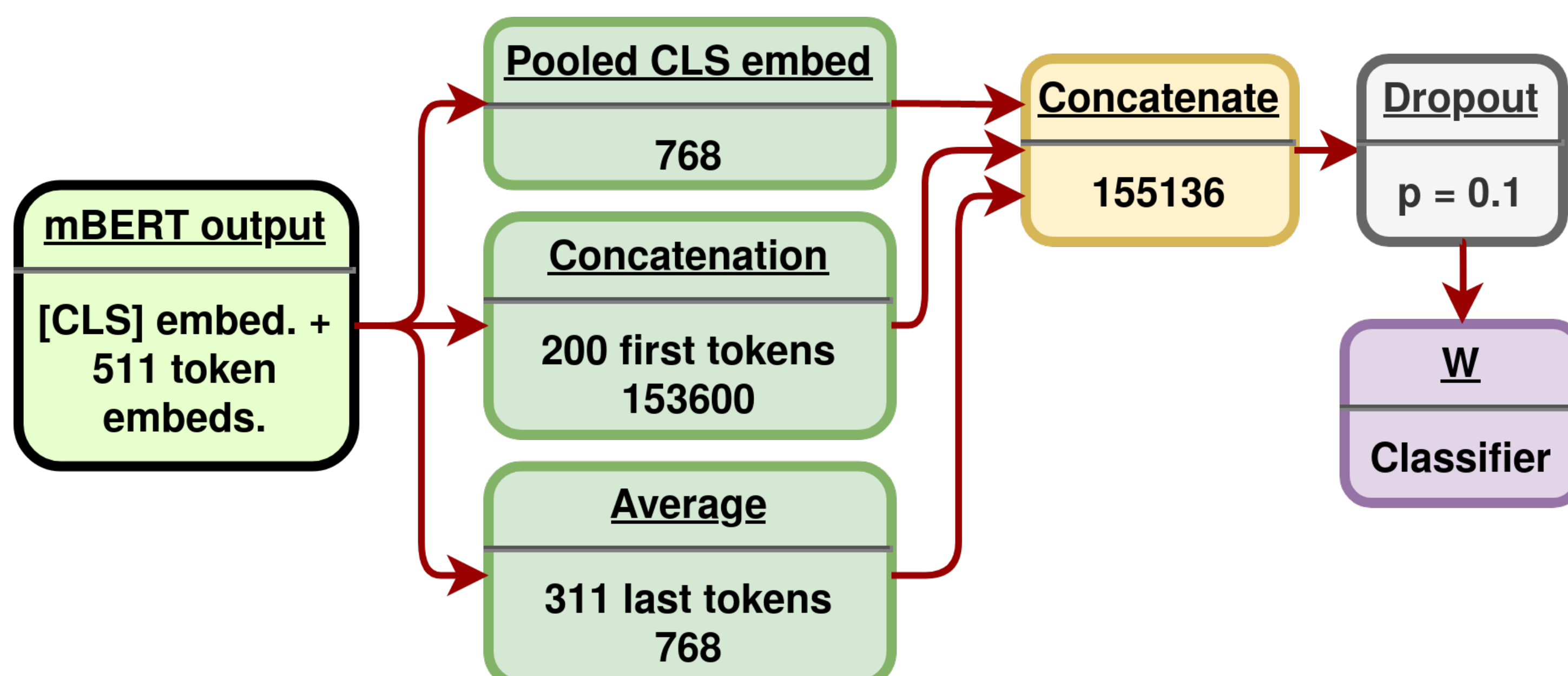


Figure 2: Linear+Concat pooling strategy and classifier.

4. Hierarchical Sequential Classification

- Explicitly leverage the ontology's hierarchical structure
- Gated Recurrent Units (GRU) layer sequentially predicts the 4 hierarchical label levels
- At each GRU step, an additional more fine-grained label is predicted
- Pooling strategy: only [CLS] token embedding

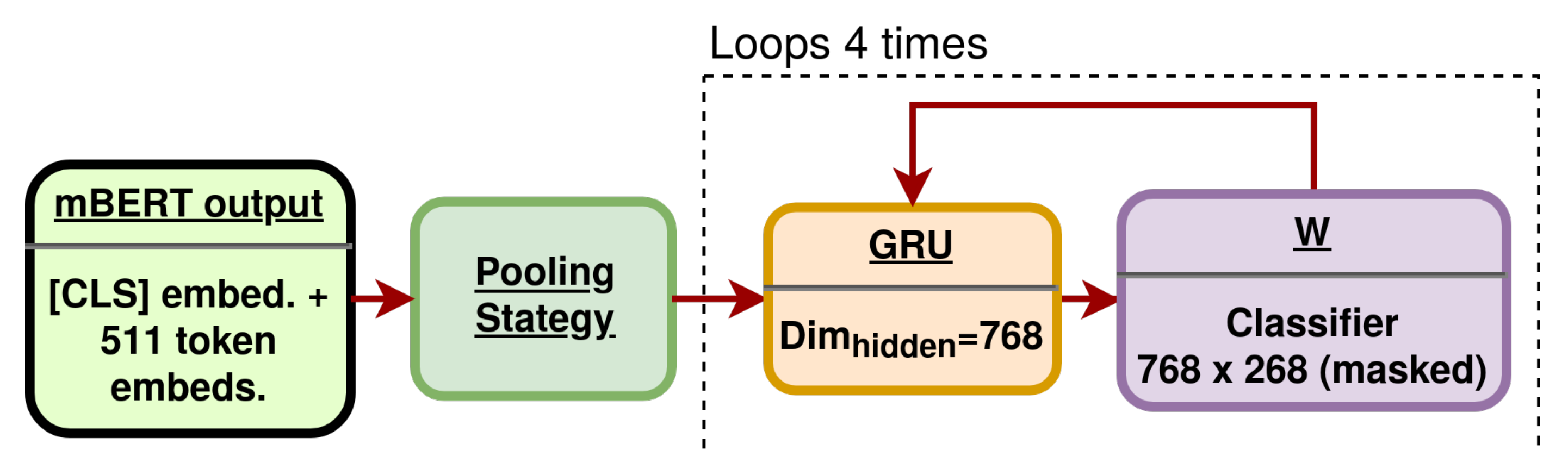


Figure 3: Hierarchical sequential classifier architecture.

5. Multi-level Hierarchical Classification

- Same architecture as Linear+Concat
- To leverage hierarchy, gold labels were decomposed into their hierarchical ancestors
- Model learns hierarchical steps that lead to leaf labels

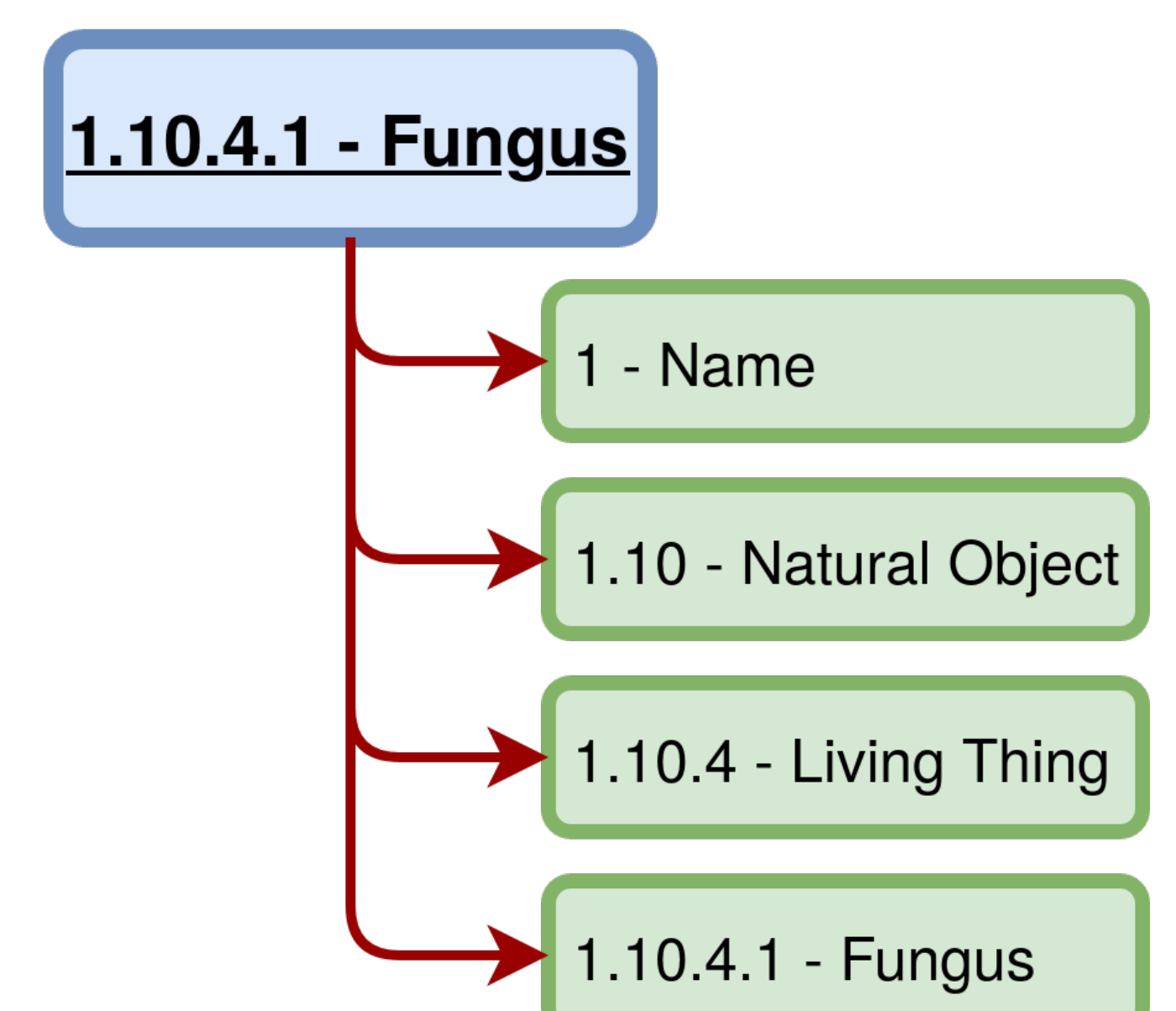


Figure 4: Label decomposition.

6. Results

- Best pooling strategy for linear classifier is Linear+Concat
- Very similar performance for Linear+Concat and GRU+CLS models
- Multi-level hierarchical model tends to perform slightly worse
- Zero-shot evaluation for Czech (CS) and Norwegian (NO) shows small decrease in performance
- Zero-shot performance with average μ F1 decrease of 3.7% for Linear+Concat and 4.1% for GRU+CLS, on leaderboard set.
- Official test set results consistent with leaderboard set.

	LINEAR+CONCAT	MULTI-LEVEL HIERARCHICAL	GRU+CLS
EN	0.739	0.713	0.707
ES	0.744	0.739	0.751
FR	0.726	0.696	0.735
DE	0.758	0.743	0.720
ZH	0.735	0.598	0.754
RU	0.745	0.723	0.730
PT	0.699	0.703	0.710
IT	0.711	0.702	0.734
AR	0.683	0.678	0.702
TR	0.732	0.711	0.699
NL	0.724	0.738	0.729
PL	0.766	0.701	0.722
KO	0.746	0.721	0.738
CS	0.692	-	0.692
NO	0.717	-	0.700

Table 1: μ F1 on leaderboard set.

	LINEAR+CONCAT	GRU+CLS
EN	0.8012	0.8127 (5th)
ES	0.8072 (5th)	0.8030
FR	0.7852 (3rd)	0.7793
DE	0.7983	0.8024 (5th)
ZH	0.7937 (3rd)	0.7838
RU	0.8308 (2nd)	0.8260
PT	0.8188	0.8236 (2nd)
IT	0.8189	0.8192 (4th)
AR	0.7545	0.7627 (1st)
TR	0.8323	0.8436 (5th)
NL	0.8126 (5th)	0.8095
PL	0.8346 (5th)	0.8273
KO	0.8104	0.8151 (5th)
CS	-	0.8119 (5th)
NO	-	0.7839 (5th)

Table 2: μ F1 on official test set.