Text Classification and Sequence Tagging

NLP

Baymurzina Dilyara

Agenda

01 Text Classification

02 Sequence Tagging

03 Classification problems

TEXT CLASSIFICATION

Text Classification Task

Given a collection of documents $\mathcal{D}$ and a set of classes $\mathcal{C}$.

The target function $\mathcal{F}^*: \mathcal{D} \rightarrow \mathcal{C}$ is unknown except of the limited training set objects $\mathcal{X}_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

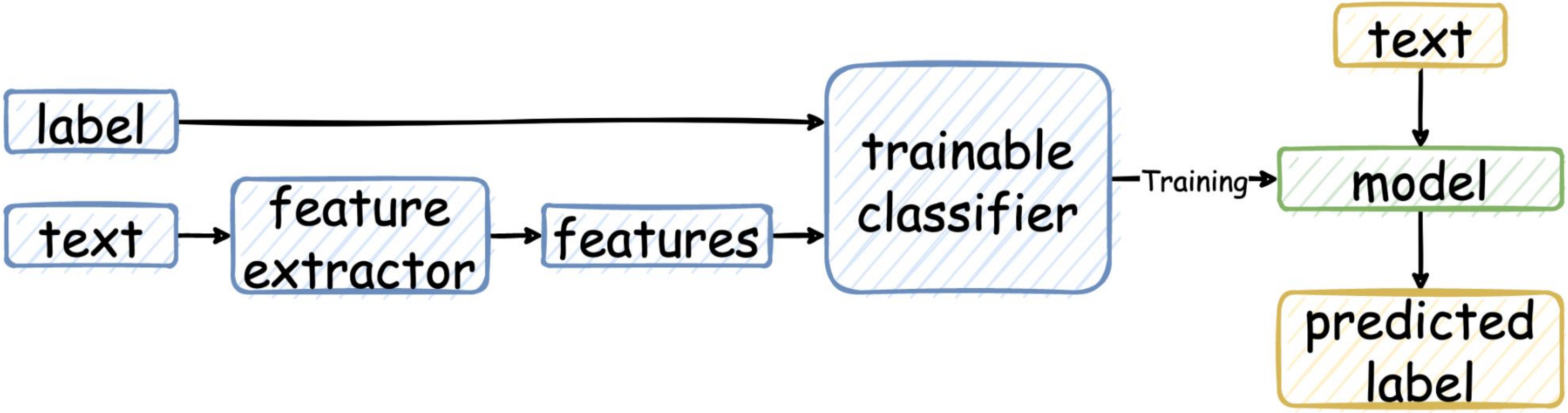
Build a model $\mathcal{F}: \mathcal{D} \rightarrow \mathcal{C}$ close to $\mathcal{F}^*$.

Example:

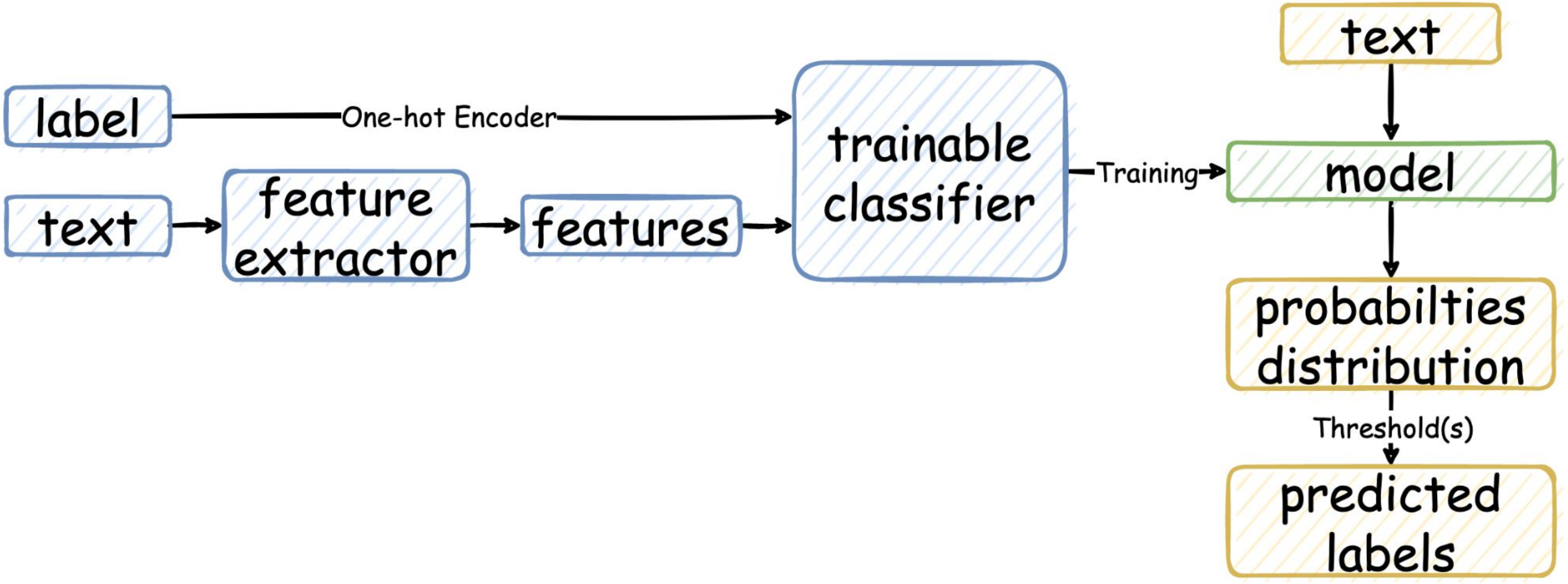
"I feel inspired at this Summer school!" → positive

"I feel asleep. Leave me alone!" → negative

Text Classification Pipeline



Text Classification Pipeline



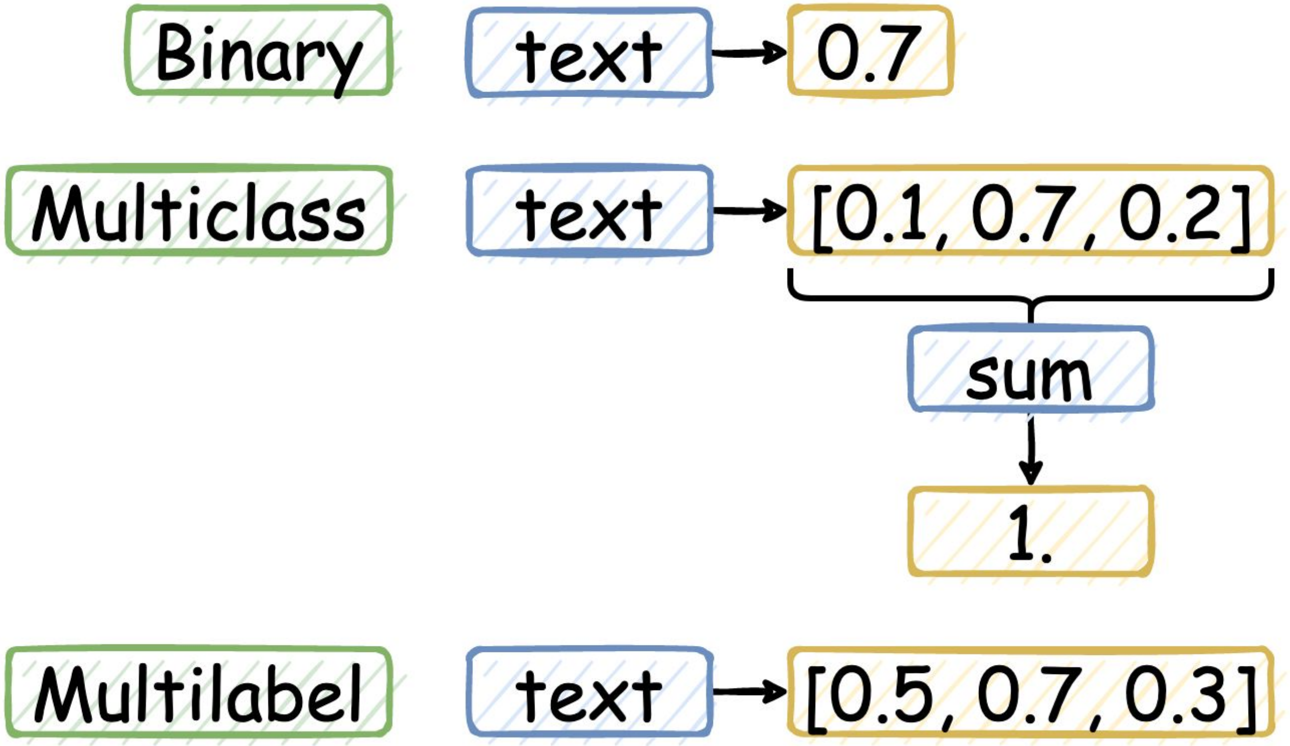
Classification Types

Binary classifier predicts probability to belong to the target class.

Multiclass classifier predicts one label of the several target classes.

Multilabel classifier predicts multiple labels per sample.

Classification Types



Application Examples

- Topics: movies, books, education, ...
- Sentiment: positive, negative, neutral, ...
- Toxicity: toxic, non-toxic, hate speech, insult, threat, ...
- Emotions: surprise, neutral, anger, joy, ...
- Dialogue Acts: open question, command, statement, ...
- Intents: greeting, yes, no, opinion request, ...
- Factoidness: factoid, non-factoid
- Frequently Asked Questions (FAQ)

Common Approaches

- Pattern-matching or dictionary-based approaches
- Machine Learning
  - mostly on full text embeddings
- Neural Networks:
  - char-level
  - token-level
  - sentence-level
  - full-text-level

Rule-based or Dictionary-based Approaches

Patterns
Dictionaries → Conditions → Hand-crafted Algorithms

`surprise` dictionary:
Wow!
Really?!
Unexpectedly!

detect(`surprise`, text):
if any word from `surprise` dictionary, assign text to `surprise`

for emotion in [`surprise`, `joy`, ...]:
detect(emotion, text)

ML-based Approaches

- Text embeddings:
  - TF-IDF, bag-of-words, word2vec, GloVe, fastText
- Algorithms:
  - LogRegression, SVC, Nearest Neighbours, Decision Trees…

Text → Vector → Model → Probabilities distribution

Neural Network Approaches

- Text is represented as a set of symbols/words from Vocabulary
- Each symbol/word is represented as a trainable vector
- Embedding Layer is trainable matrix
- No info for OOV (out-of-vocabulary) words

Text → Trainable Embedding Matrix of Limited Size → Model → Probabilities distribution

Trainable Embeddings for Limited Vocabulary

```
>>> model = tf.keras.Sequential()
>>> model.add(tf.keras.layers.Embedding(1000, 64, input_length=10))
>>> # The model will take as input an integer matrix of size (batch,
>>> # input_length), and the largest integer (i.e. word index) in the input
>>> # should be no larger than 999 (vocabulary size).
>>> # Now model.output_shape is (None, 10, 64), where `None` is the batch
>>> # dimension.
>>> input_array = np.random.randint(1000, size=(32, 10))
>>> model.compile('rmsprop', 'mse')
>>> output_array = model.predict(input_array)
>>> print(output_array.shape)
(32, 10, 64)
```

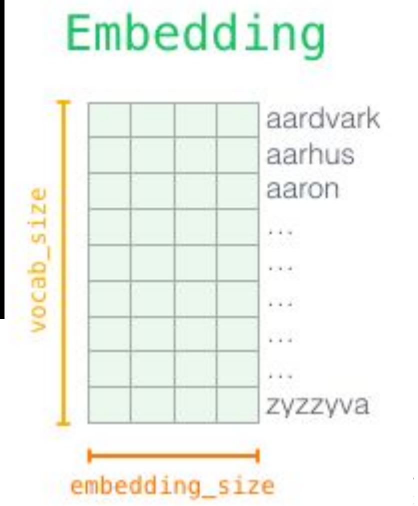


Image source

Neural Network Approaches

fastText vector representations are the state-of-the-art trainable text embeddings in 2017.

Text → Vector → Model → Probabilities distribution

Image source

Neural Network Approaches

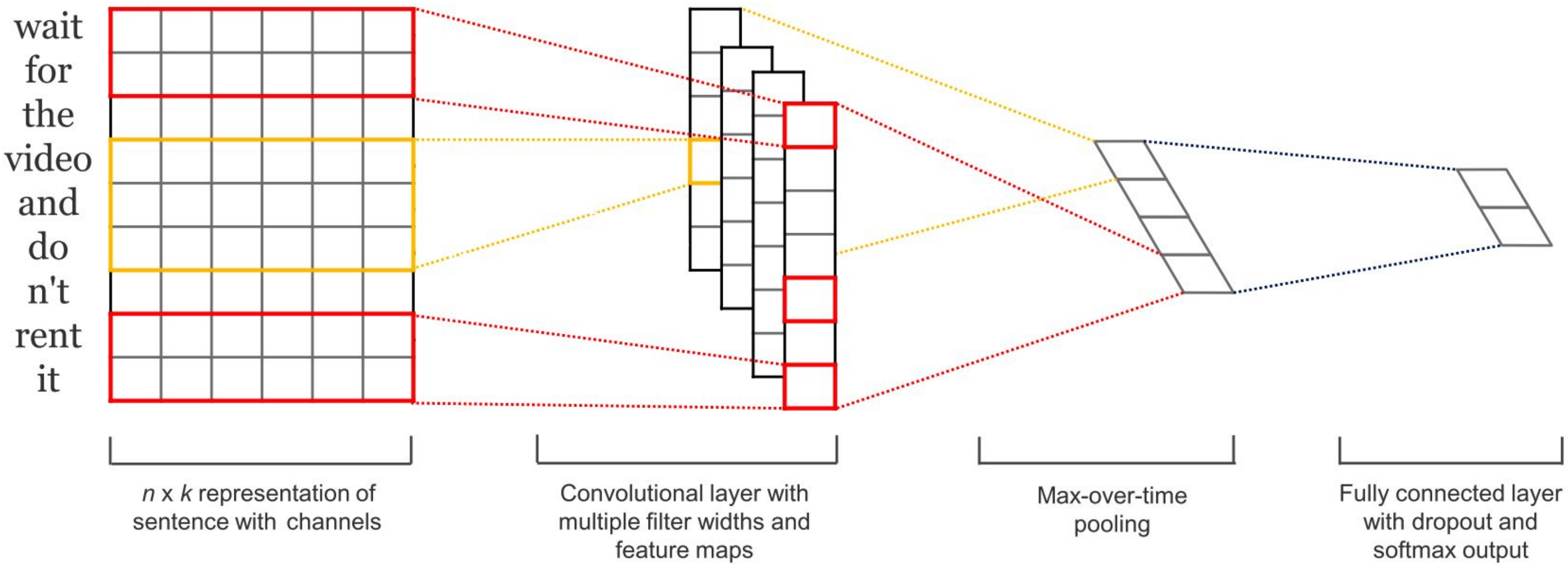


Image source

ELMo

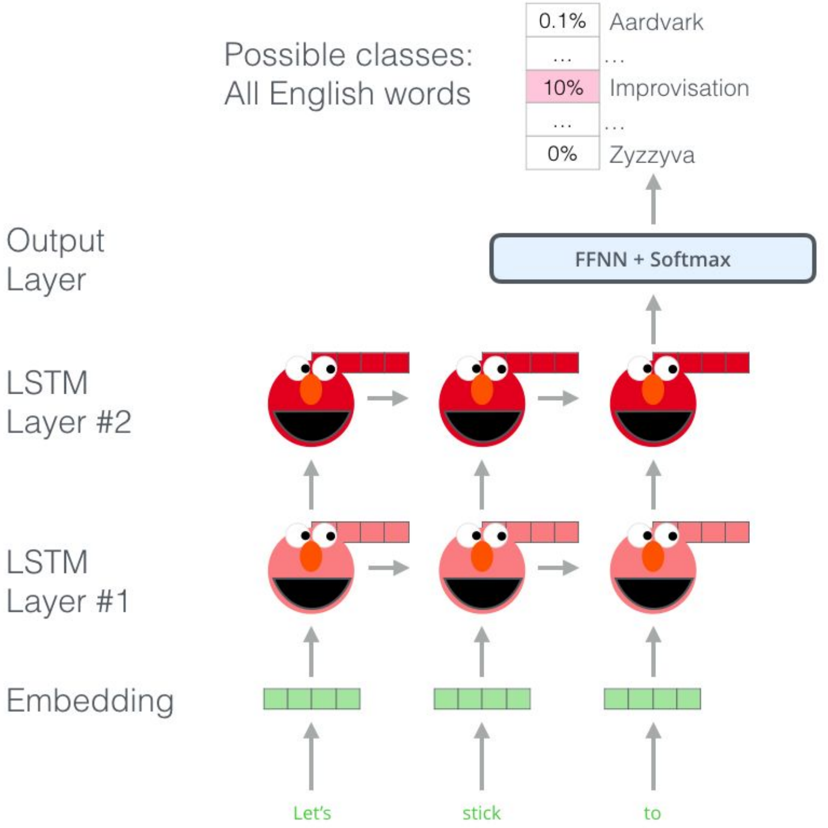
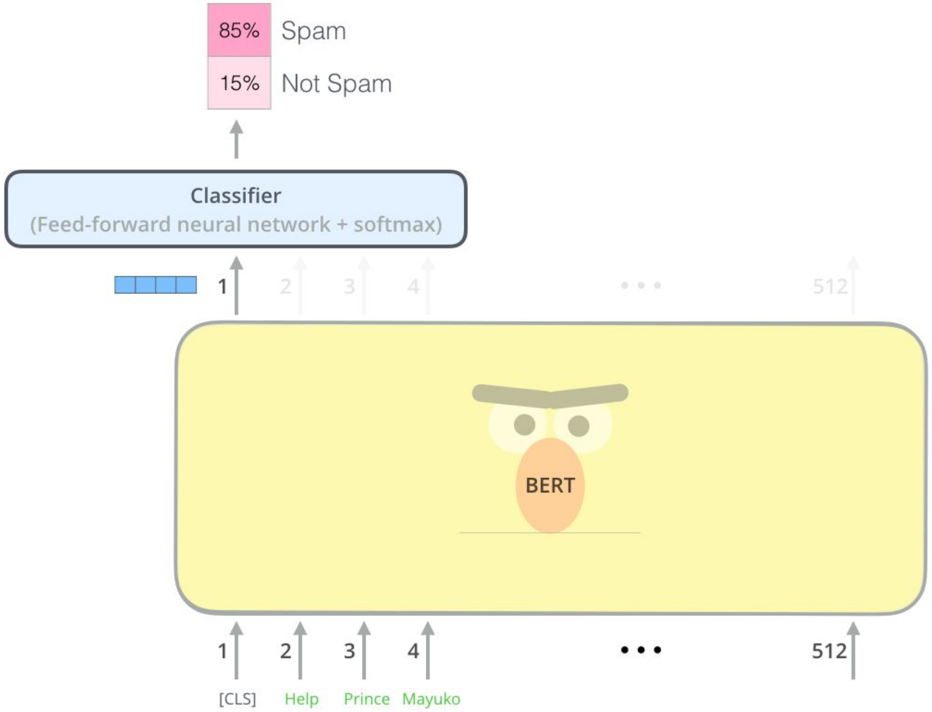


Image source

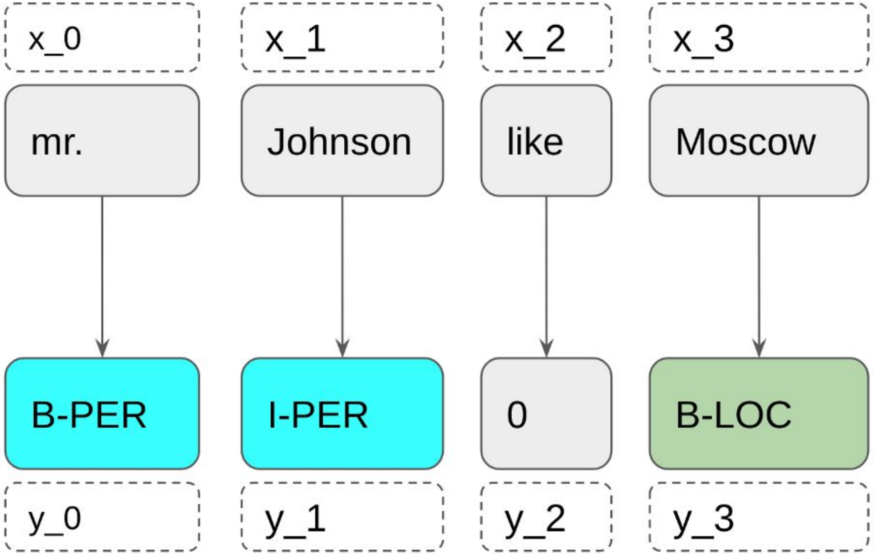
BERT


85% Spam
15% Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

SEQUENCE TAGGING

Sequence Tagging Task

Given a sequence $\boldsymbol{x}=\{x_0, ..., x_m\}$, assign label for each element of $\boldsymbol{x}$.



Application Examples

- NER - Named Entity Recognition
- PoS-tagging (part-of-speech)
- Morpho-tagging
- Entity Recognition

BIO-notation:

B - beginning,

I - intermediate,

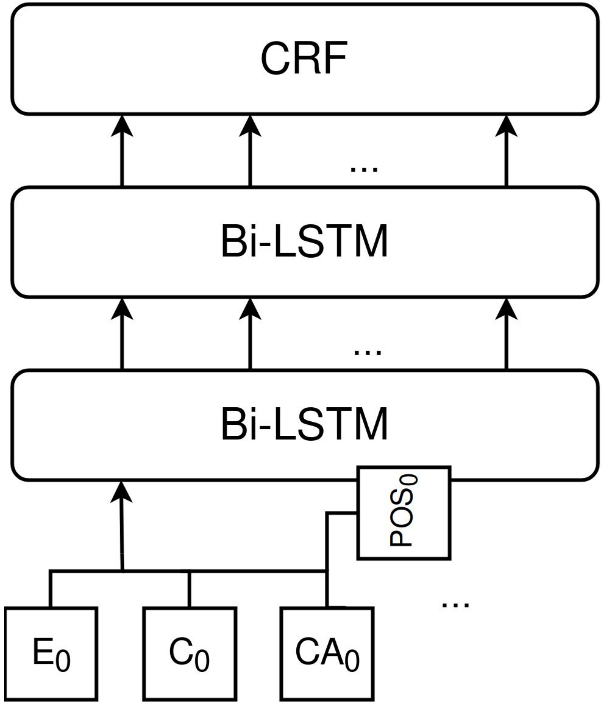
O - outside

```
The United States of    America
O    B-LOC  I-LOC  I-LOC  I-LOC

has an intelligent leader in D.C.
 O  O       O         O    O  B-LOC

, Dick  Cheney of Halliburton .
O B-PER I-PER  O   B-ORG      O
```

AIRI × Университет Сириус

Neural Network Approaches

Bi-LSTM + CRF + Char + Capitalization + POS



Arkhipov, et al. 2018

ELMo for Sequence Tagging

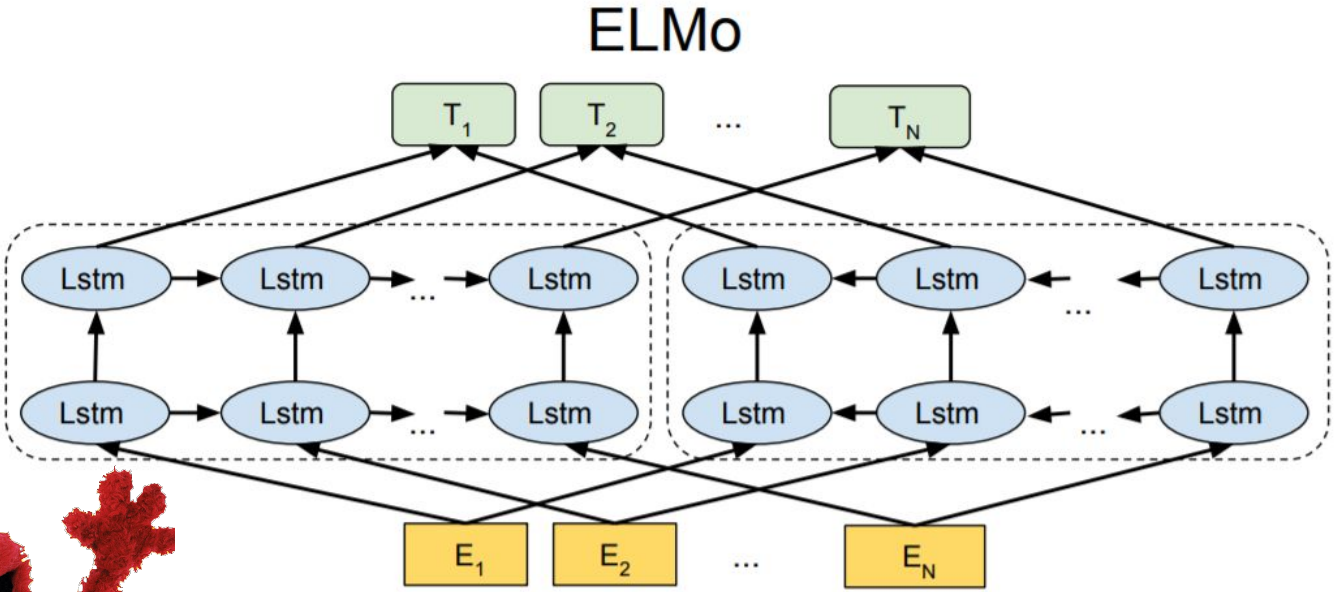


Image source

BERT for Sequence Tagging

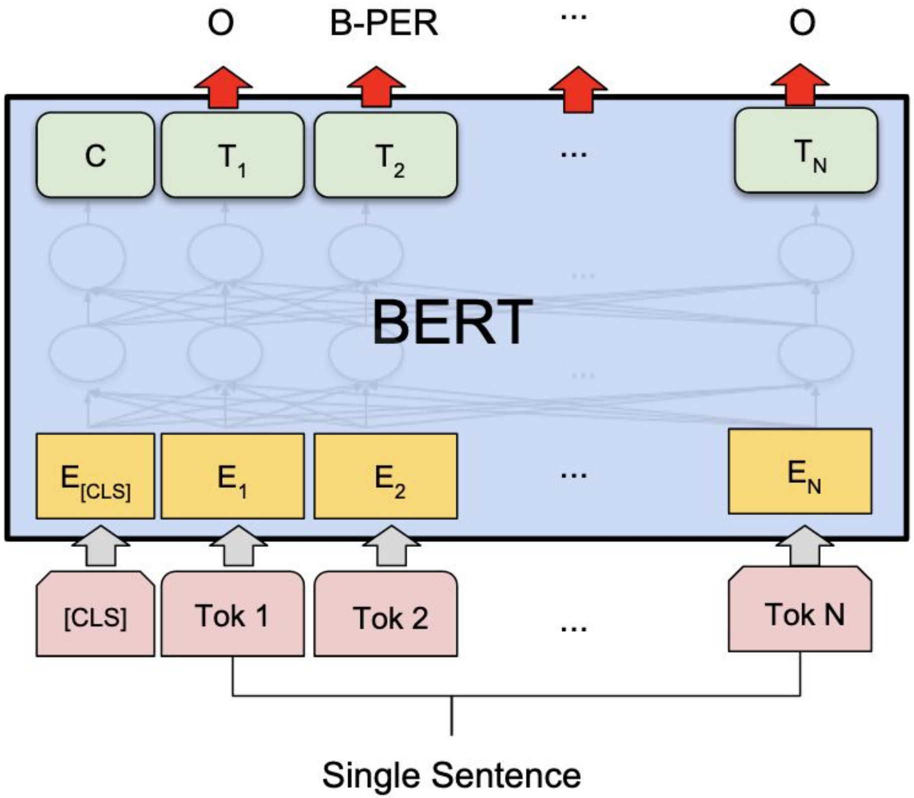
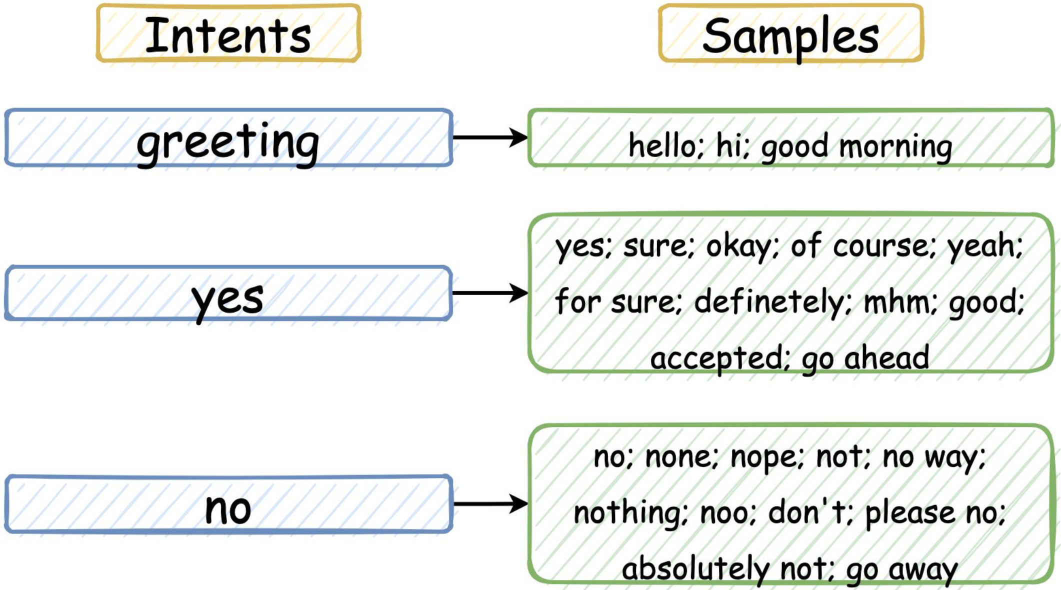


Image source

CLASSIFICATION PROBLEMS

Imbalanced Classes

For real-life classification datasets, it is common to have a very different number of samples per class.



Imbalanced Classes

- Downsampling
- Weighting classes in a loss function
- Data Augmentation

Insufficient Number of Samples

- Data Augmentation
- Learnable unseen detectors
- Intent prototypes within external knowledge
- Leveraging common sense knowledge graphs
- Utilizing class description and reformulating as NLI task
- Highly informative class labels as a second input to model

Data Augmentation

Text Attack:

- replacing words with WordNet synonyms
- replacing words with neighbors in the counter-fitted embedding space
- substituting, deleting, inserting, and swapping adjacent characters
- combination of word insertions, substitutions and deletions
- contraction/extension and substituting names, locations, numbers
- replacing, inserting, and merging with a pre-trained MLM

Pre-training Intent Task

- Pre-train BERT-classifier on 1k samples for intent classification;
- Fine-tune on the domain for MLM task;
- Extract features using the fixed IntentBERT;
- Apply simple classifier to the features.

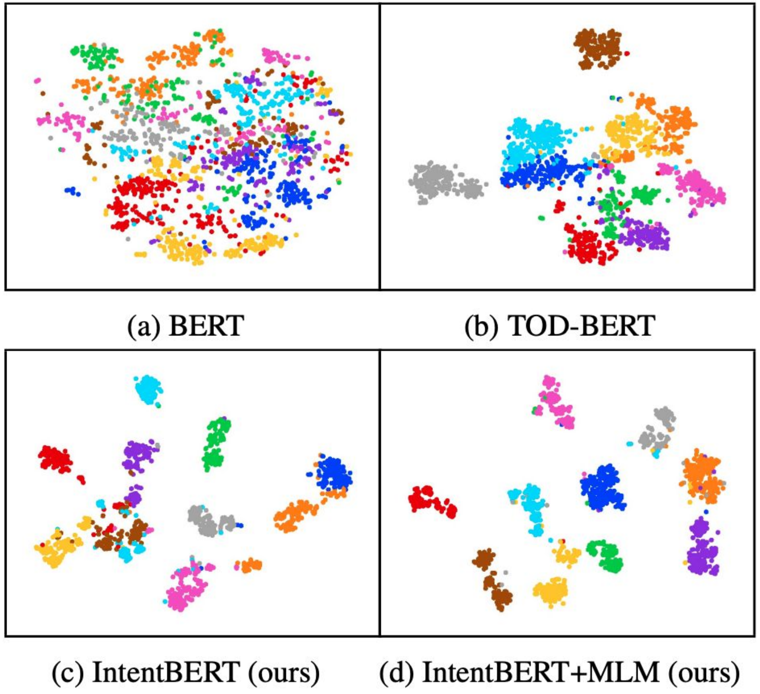
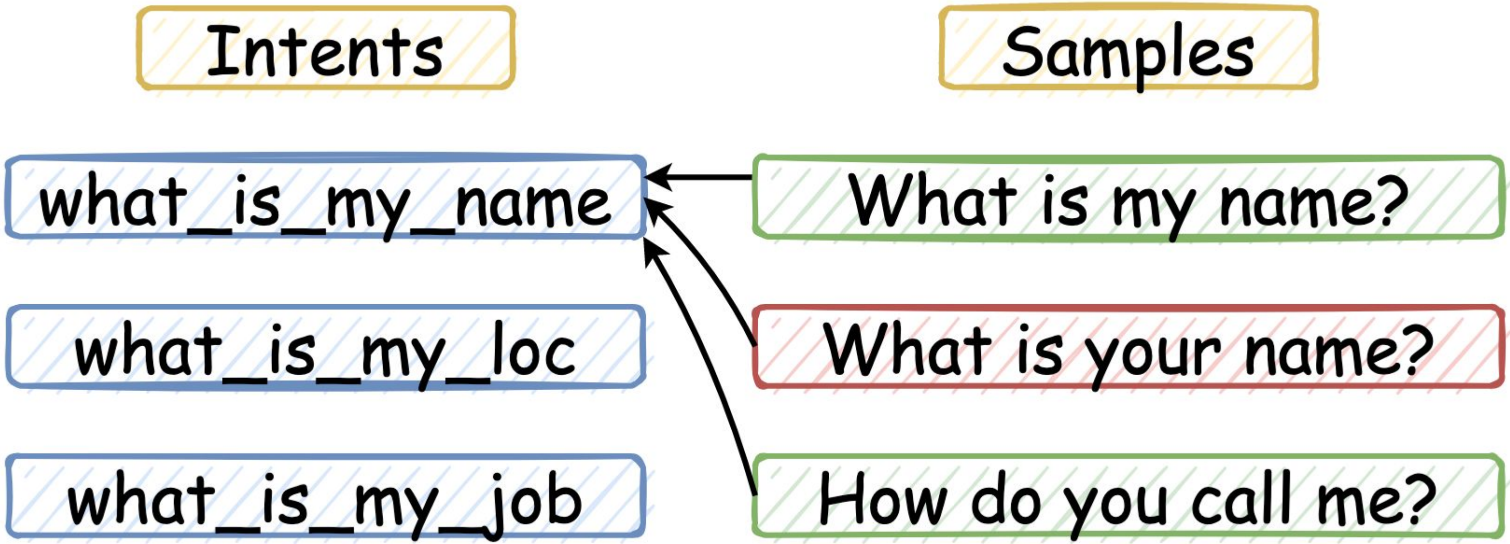


Figure 1: Visualization of the embedding spaces with t-SNE. We randomly sample 10 classes and 500 data per class from BANKING77 (best viewed in color).

Paper

Out-of-Scope Problem

For multilabel classification, it is common to get a lot of false positive labels.



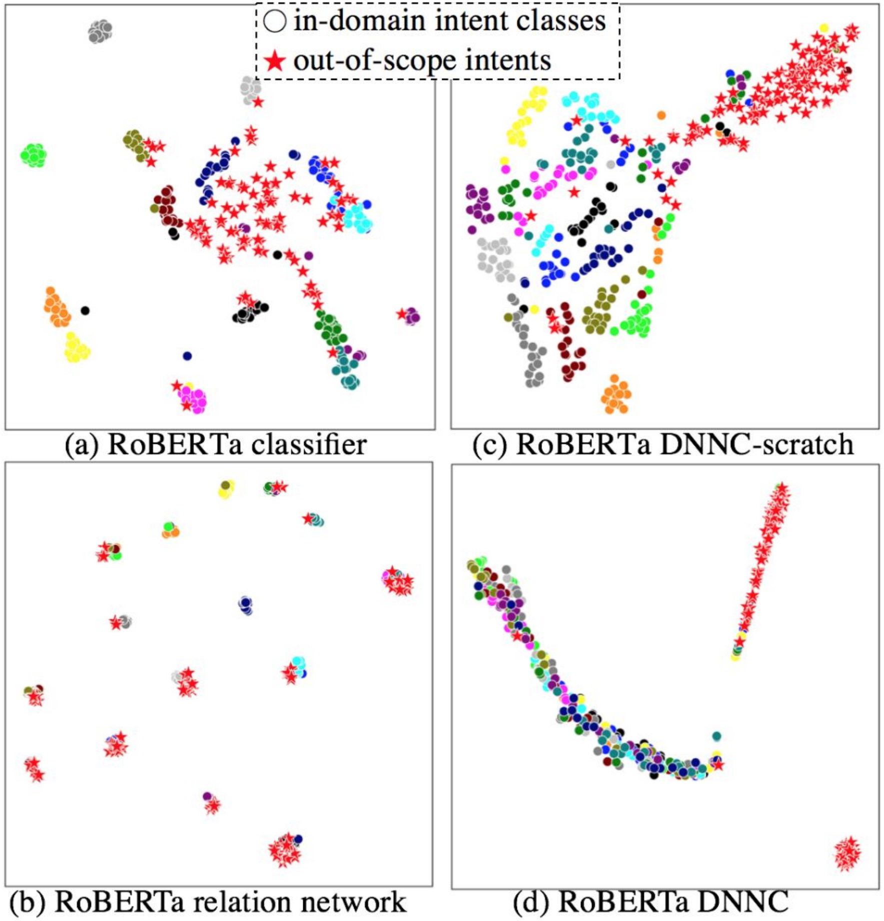
AIRI × Университет Сириус

Out-of-Scope Problem

- Add negative (not belonging to all classes) samples by hands
- 2-stage approach:
  - binary classification whether a text belongs to the considered classes;
  - multiclass classification among the considered classes.
- Classifier with additional class for out-of-scope samples
- Classifier with threshold(s) for in-scope classes
- One-Vs-The-Rest classes setup - DNNC approach
  - Binary classifier for each label, negative samples are all samples for other classes

DNNC

- Form positive (texts of the same intent) and negative (texts of different intents) examples;
- Utilize BERT pair-wise encoding;
- Utilize entailment-NLI pre-trained model to train model which is close to 1.0 when samples from the same class, and close to 0.0 otherwise.

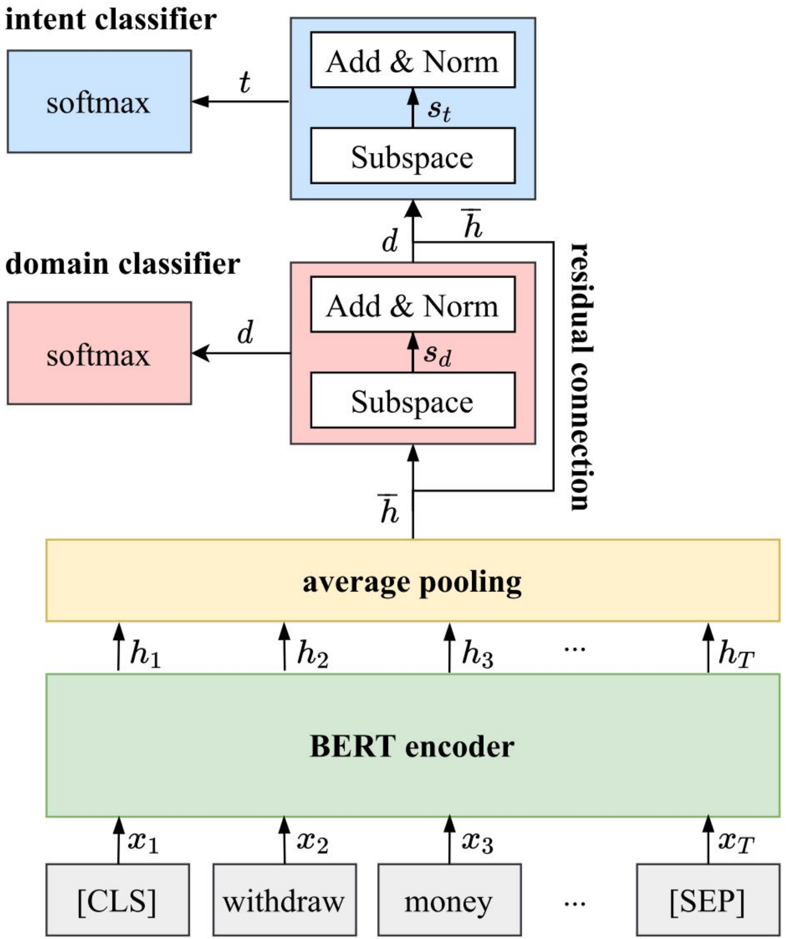


Paper

Multitask to improve classification quality

BERT-Joint architecture

- Combine domain detection and intent classification

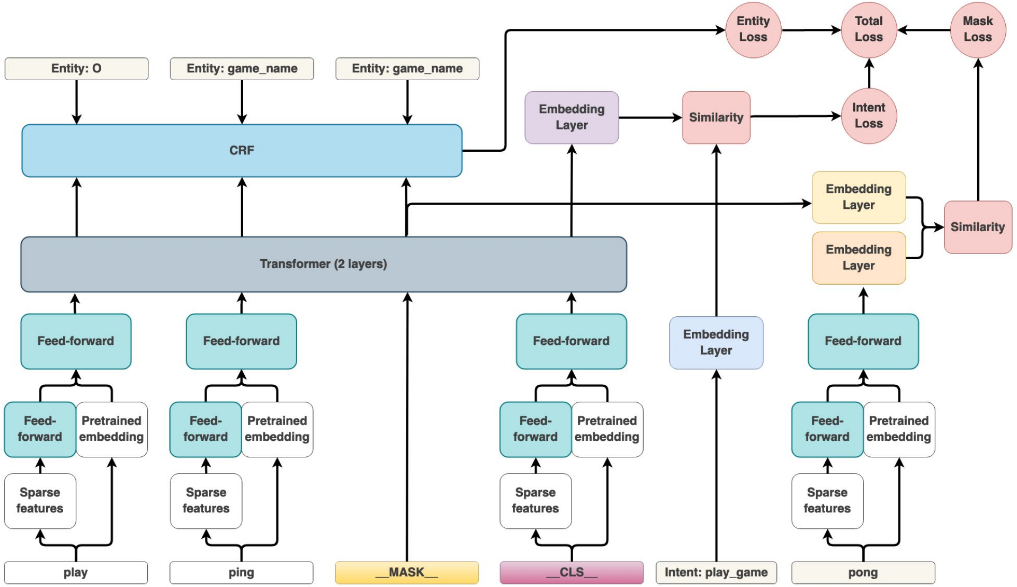


Paper

Multitask to improve classification quality

DIET – Dual Intent and Entity Transformer:

- Combine intent classification and entity recognition tasks.
- Combine pre-trained embeddings with word- & char-level ngrams.



Paper

AIRI × Университет Сириус

Artificial Intelligence Research Institute

airi.net

airi_research_institute

AIRI Institute

AIRI Institute

AIRI_inst

artificial-intelligence-research-institute