MULTI-INSTANCE LEARNING FOR RHETORIC STRUCTURE PARSING

Sergey Volkov
volkserg1@gmail.com
Devyatkin D. A.
Shvets A. V.

Outline

- Introduction in RST
- Datasets and tools
- Data pre-processing
- Models description
- Results
- Conclusion

Text processing steps

During processing, the text is analyzed at following linguistic levels:

- graphematic
- morphological;
- **syntactic**;
- semantic;

Rhetorical Structure Theory (W. Mann, S. Thompson)

Rhetorical structure theory (RST) is a theory of text organization that describes relations that hold between parts of text.

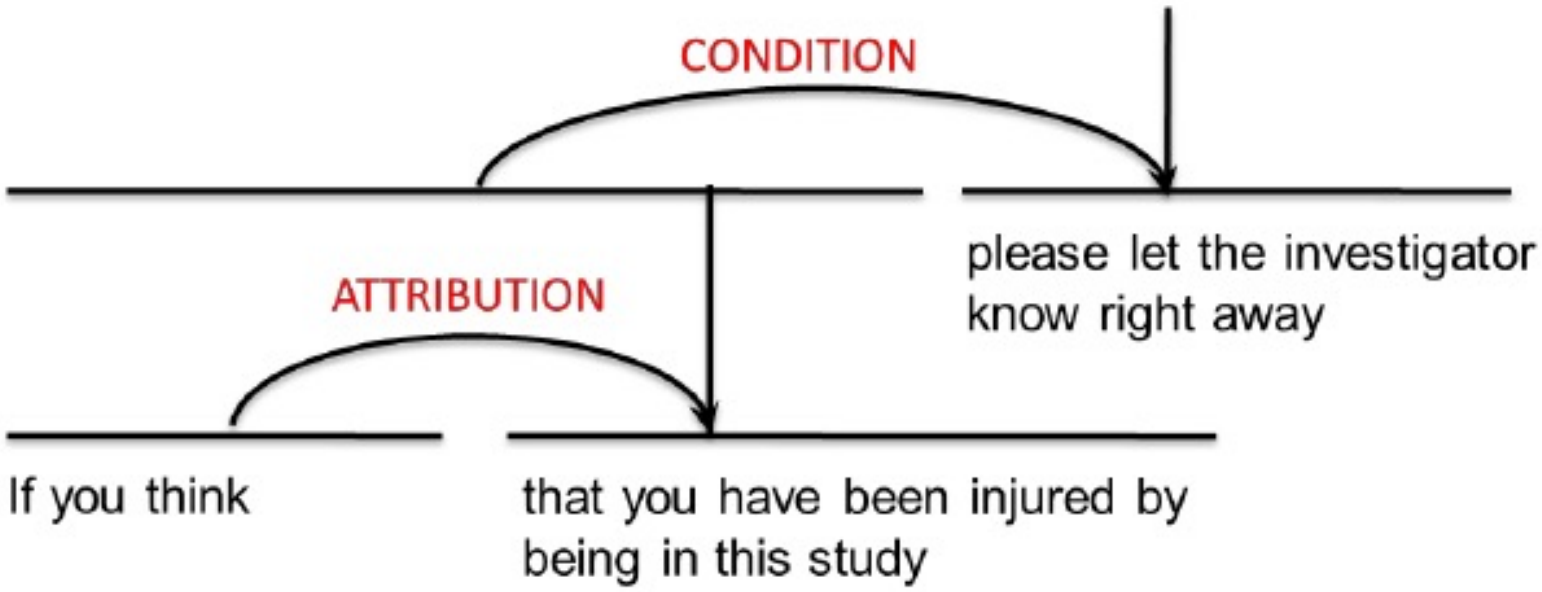


Fig.1 – RST tree, representing the rhetorical structure of text, leaves represent elementary discourse units (EDUs), arrows point from satellite to nucleus, and labels above arrows represents discourse relations.

Datasets

- **Argument Annotated Essays (AAE) corpus of Stab and Gurevych (2014a)** **(90 persuasive student essays (72 for training, 18 for testing))**
  - Using Argument Mining to Assess the Argumentation Quality of Essays (https://www.aclweb.org/anthology/C16-1158.pdf)
  - Unsupervised Learning of Discourse-Aware Text Representation for Essay Scoring (https://www.aclweb.org/anthology/P19-2053.pdf)
- **Penn Discourse Treebank Version 2.0 LDC2008T05**
  - Improving Implicit Discourse Relation Classification by Modeling Inter-dependencies of Discourse Units in a Paragraph (https://www.aclweb.org/anthology/N18-1013.pdf)
- **Penn Discourse Treebank Version 3.0 LDC2019T05**
  - Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks (https://www.aclweb.org/anthology/L16-1165.pdf) (PDTB, RST, SDRT and CCR)
- **RST Discourse Treebank LDC2002T07**
  - Cross-lingual RST Discourse Parsing (https://www.aclweb.org/anthology/E17-1028.pdf)
- **RST Signalling Corpus**
  - Can Discourse Relations be Identified Incrementally? (https://www.aclweb.org/anthology/I17-2027.pdf)

Transfer Learning

- Penn Discourse Treebank (PDTB) 2.0 и 3.0, TED Multilingual Discourse Bank

TED-MDB Zeyrek et al. 2018

- LASER + Feedforward NN +TED

Zeyu Dai and Ruihong Huang. 2018, Kurfalı M., Östling R 2019.

- BiLSTM, BERT+rule-base discourse

Nie A. et al 2019.

✓ Discourse classification to fit vector representations of sentences

✓ Better learning rate

Model	All		Books 8		Books 5	
	F1	Acc	F1	Acc	F1	Acc
GloVe-bow	17.1	41.8	27.6	47.3	41.7	52.5
Ngram-bow	28.1	51.8	44.0	58.1	54.1	63.3
BiLSTM	47.2	67.5	64.4	73.5	72.1	77.3
BERT	60.1	77.5	76.2	82.9	82.6	86.1

Main tasks

- Development of discourse relations classifier
- Using multiple datasets for better model training
  - Preparing additional automatically marked dataset
  - Implementation of machine learning methods on noisy data.
- Further application of the classifier model as a module in the analysis of the emotional charge of the text.

Data and tools

- Training set:
  - Reddit 2003-2018 (~16M pairs).
  - Pairs of all connected DU are created.
- Tuning + test:
  - Discourse Treebank LDC2002T07 (~14K+2K pairs)
- Features:
  - StanfordNLP (tokenization, morphology, syntax): https://nlp.stanford.edu/software
  - Vector representation (Word2Vec, CommonCrawl, dict size - 2.2M)
  - BERT-tokenization
- Automatic text markup:
  - *Wang Y., Li S., Wang H.* Two-stage parser: https://github.com/yizhongw/StageDP
  - *Heilman M., Sagae K*. Fast rhetorical structure theory discourse parser: https://github.com/EducationalTestingService/discourse-parsing

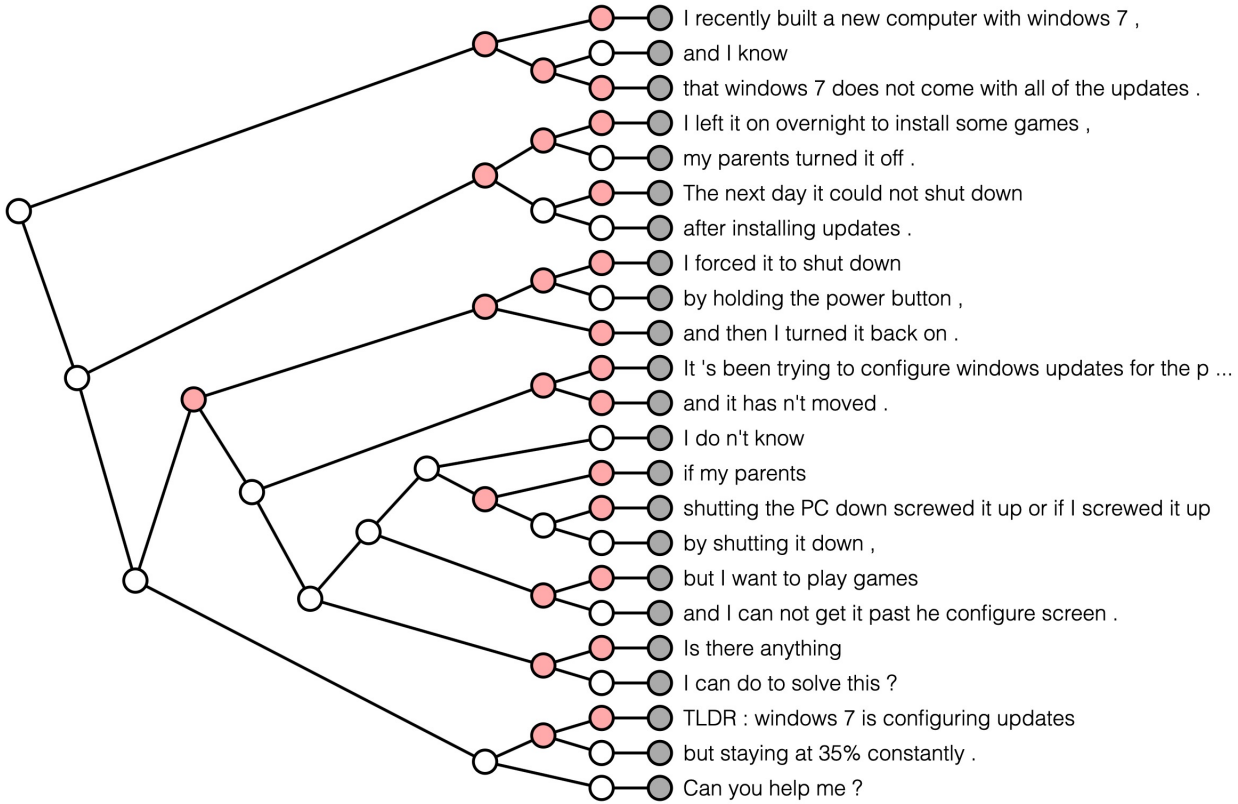
Data pre-processing

Selecting relationships using automatic text markup tools. (Two-stage parser/Fast rhetorical structure theory discourse parser).

Tokenization of selected discursive units (StanfordNLP/BERT)

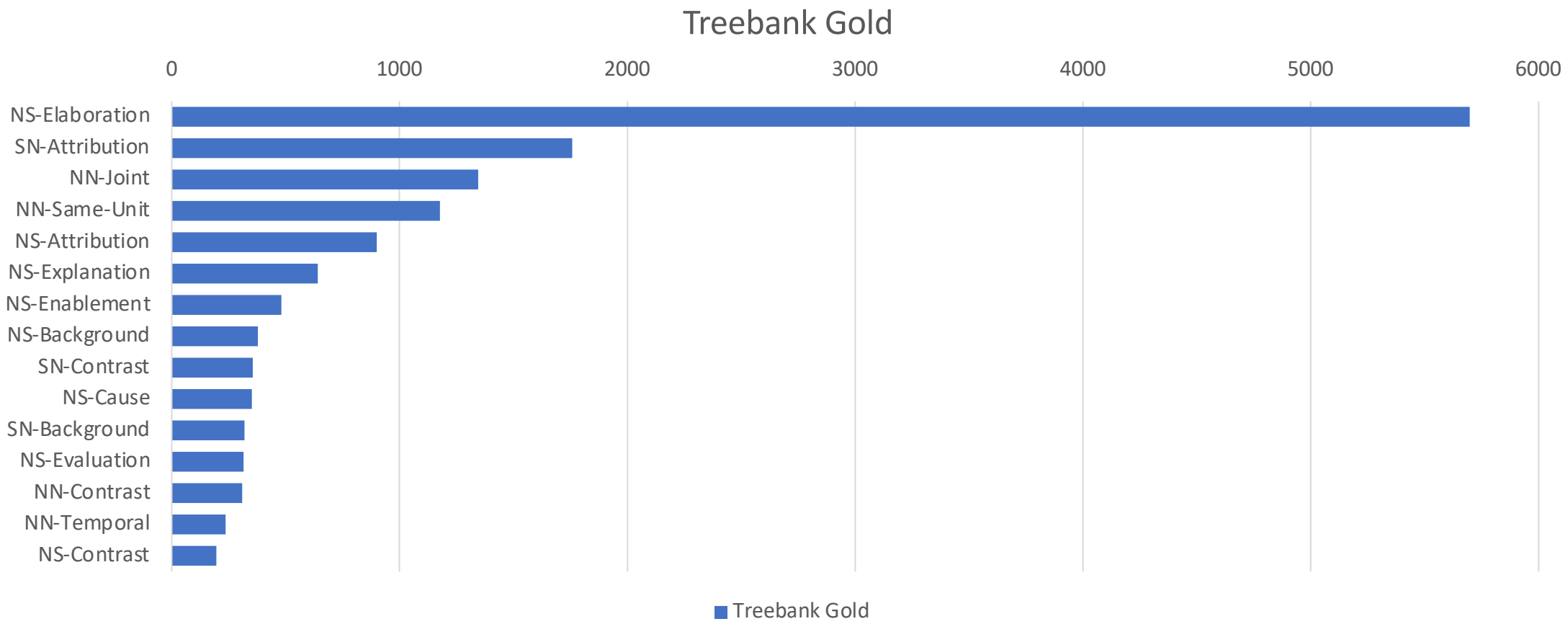
Converting a list of tokens to a list of vectors (Word2Vec)

Example of a marker result. Discourse parser (*Heilman M)*



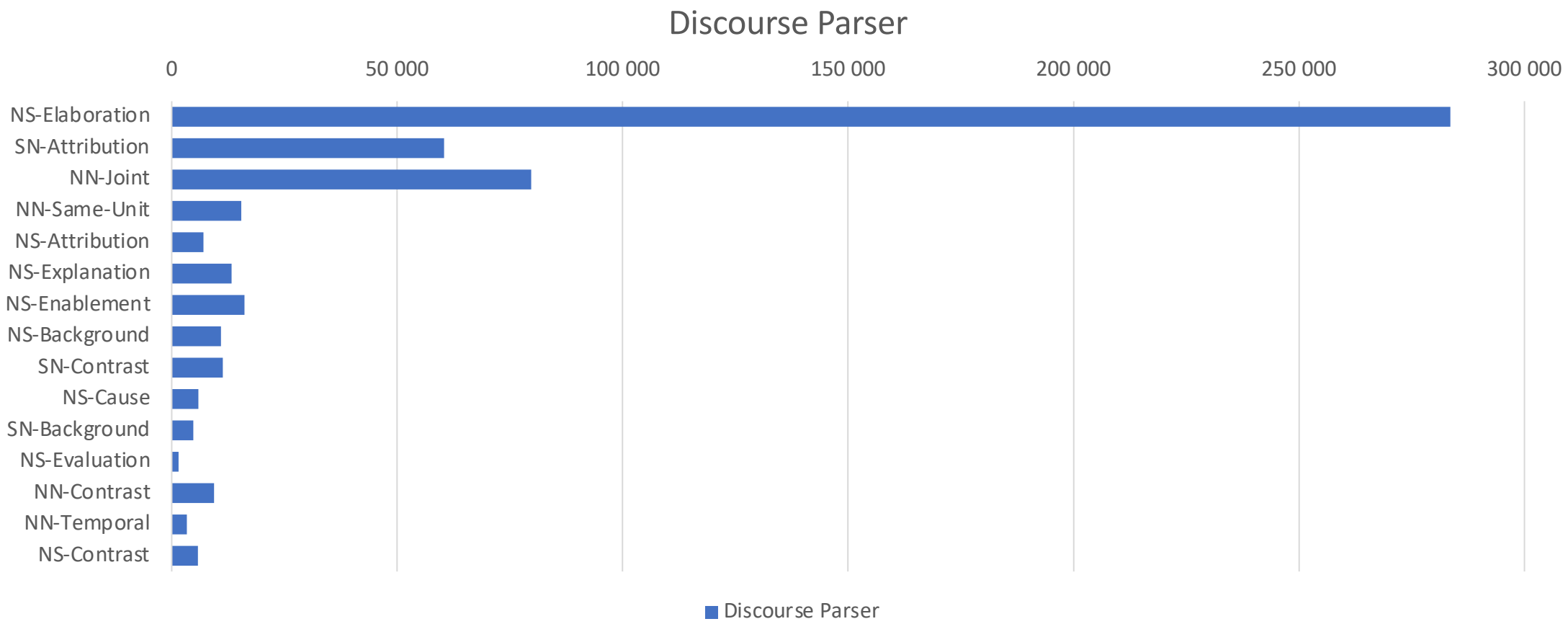
Part 1	Part 2	Relation
I recently built a new computer with windows 7	and I know that windows 7 does not come with all of the updates .	NN-Joint
and I know	that windows 7 does not come with all of the updates	SN-Attribution
I forced it to shut down	by holding the power button	NS-Manner-Means
I forced it to shut down by holding the power button	and then I turned it back on	NN-Temporal

Class balance (top-15 of gold treebank)



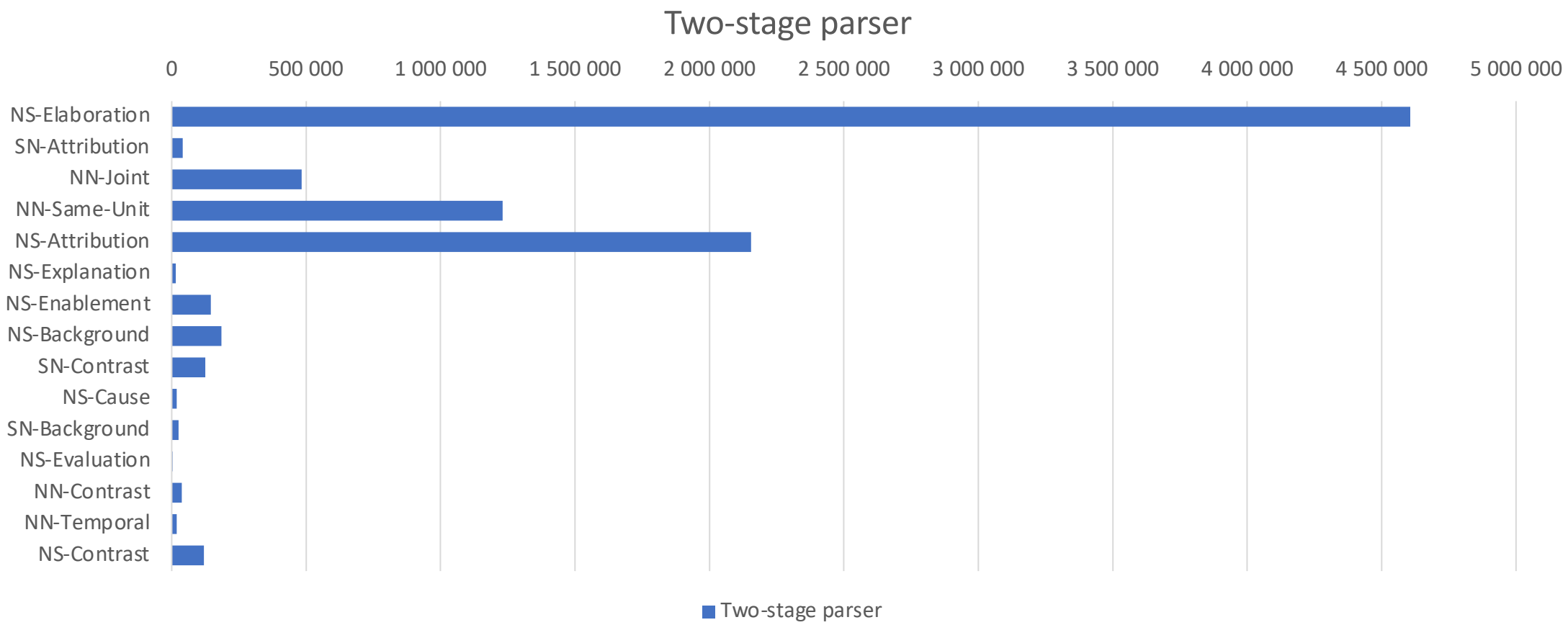
Total: 15 815 (Discourse Treebank LDC2002T07)

Class balance (top-15 of gold treebank)



Total: 561 382 (Reddit 2003-1018)

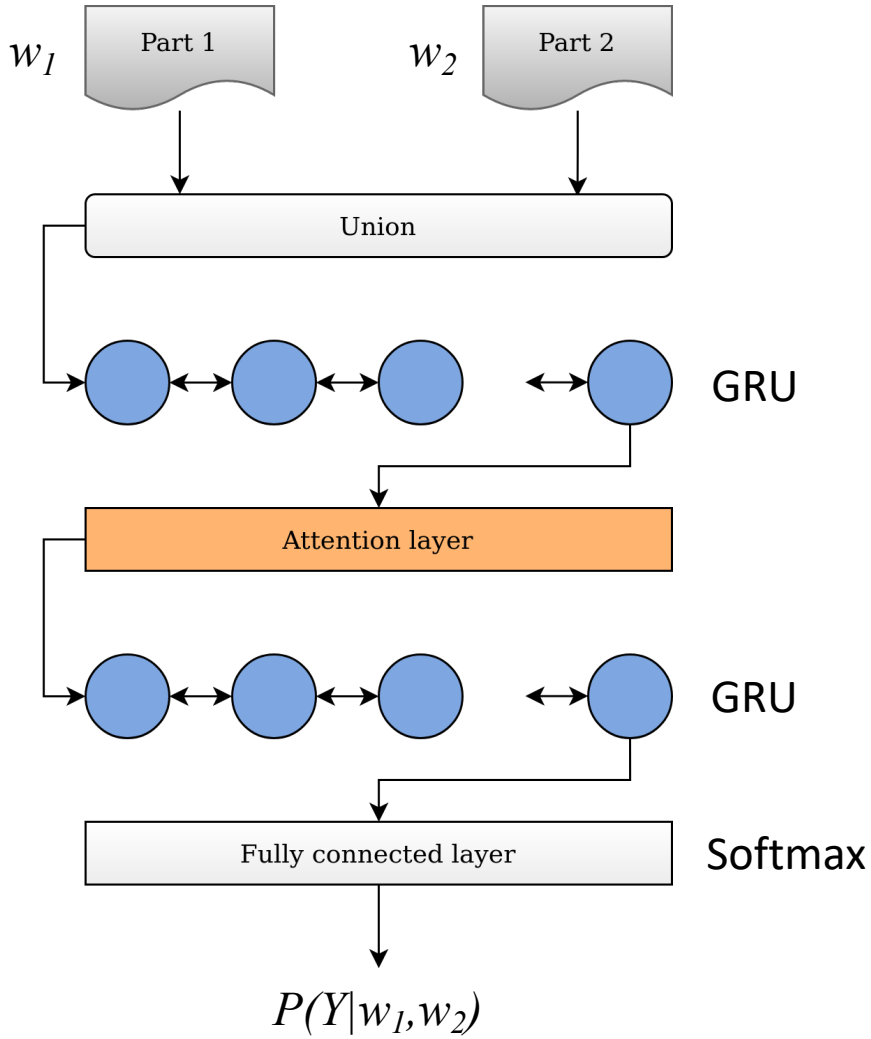
Class balance (top-15 of gold treebank)



Total : 15 616 714 (Reddit 2003-1018)

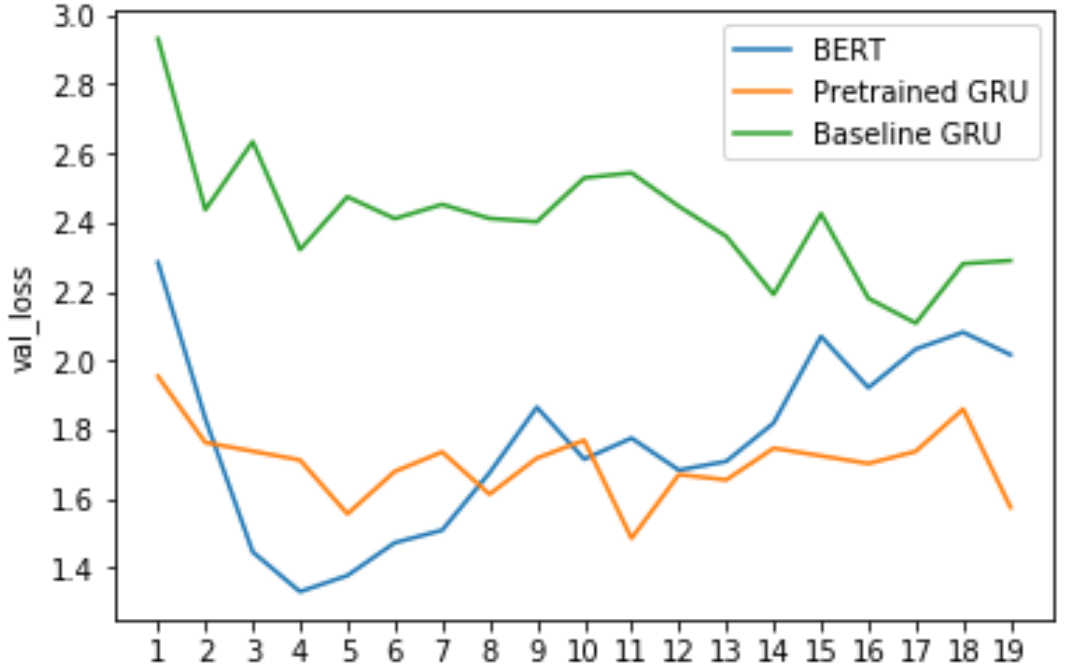
Discourse relations classification model

Bi-GRU, 1,030,892 parameters:



- Bi-GRU Pretrained – train on Reddit + tuning on Discourse Treebank
- Bi-GRU Baseline – train on Discourse Treebank
- BERT – tuning on Discourse Treebank

Validation loss:



Results (top classes by F1)

BERT architecture:

- Input layer (512 neurons)
- BERT layer – 108 891 648 parameters
- Fully connected layer – 256 neurons
- Fully connected layer (softmax) – 43 neurons.

BI-GRU architecture:

- Input layers (250+250 neurons)
- BI-GRU layer – 256 neurons
- Self-attention layer – attention_width=256
- BI-GRU layer – 128 neurons
- Fully connected layer – 64 neurons
- Fully connected layer (softmax) – 43 neurons.

BERT

F_1	Класс
0.87	NS-Attribution
0.82	SN-Attribution
0.73	SN-Condition
0.63	NN-Same-Unit
0.54	NN-Joint

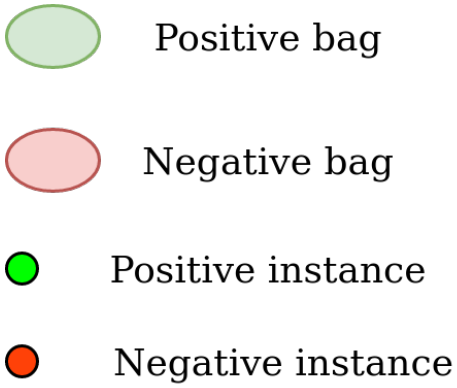
Bi-GRU (Pretrained)

F_1	Класс
0.71	NS-Attribution
0.65	SN-Attribution
0.60	NS-Elaboration
0.60	SN-Condition
0.48	NS-Enablement

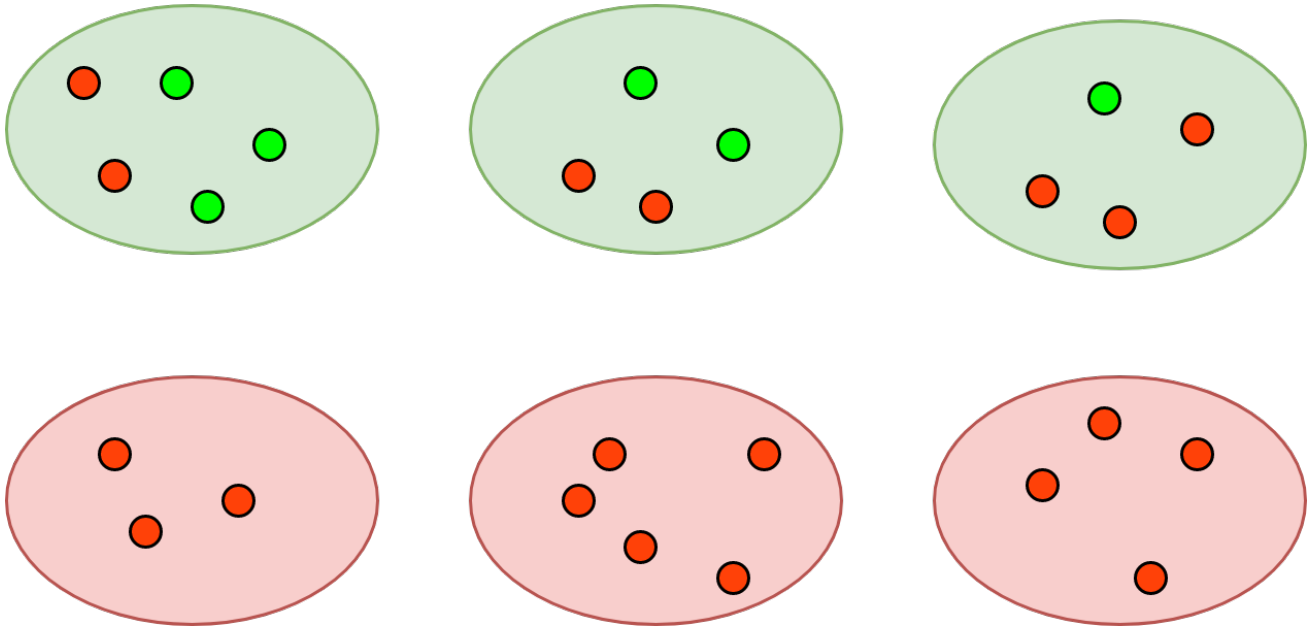
Bi-GRU (Baseline)

F_1	Класс
0.59	NS-Attribution
0.58	SN-Attribution
0.20	NS-Elaboration
<0.05	SN-Condition
<0.05	NS-Enablement

Multi Instance Learning (MIL)



- **Multi Instance Learning** – is a type of supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances.
- In the simple case of a binary classification with multiple instances, a bag can be marked negative if all instances in it are negative.

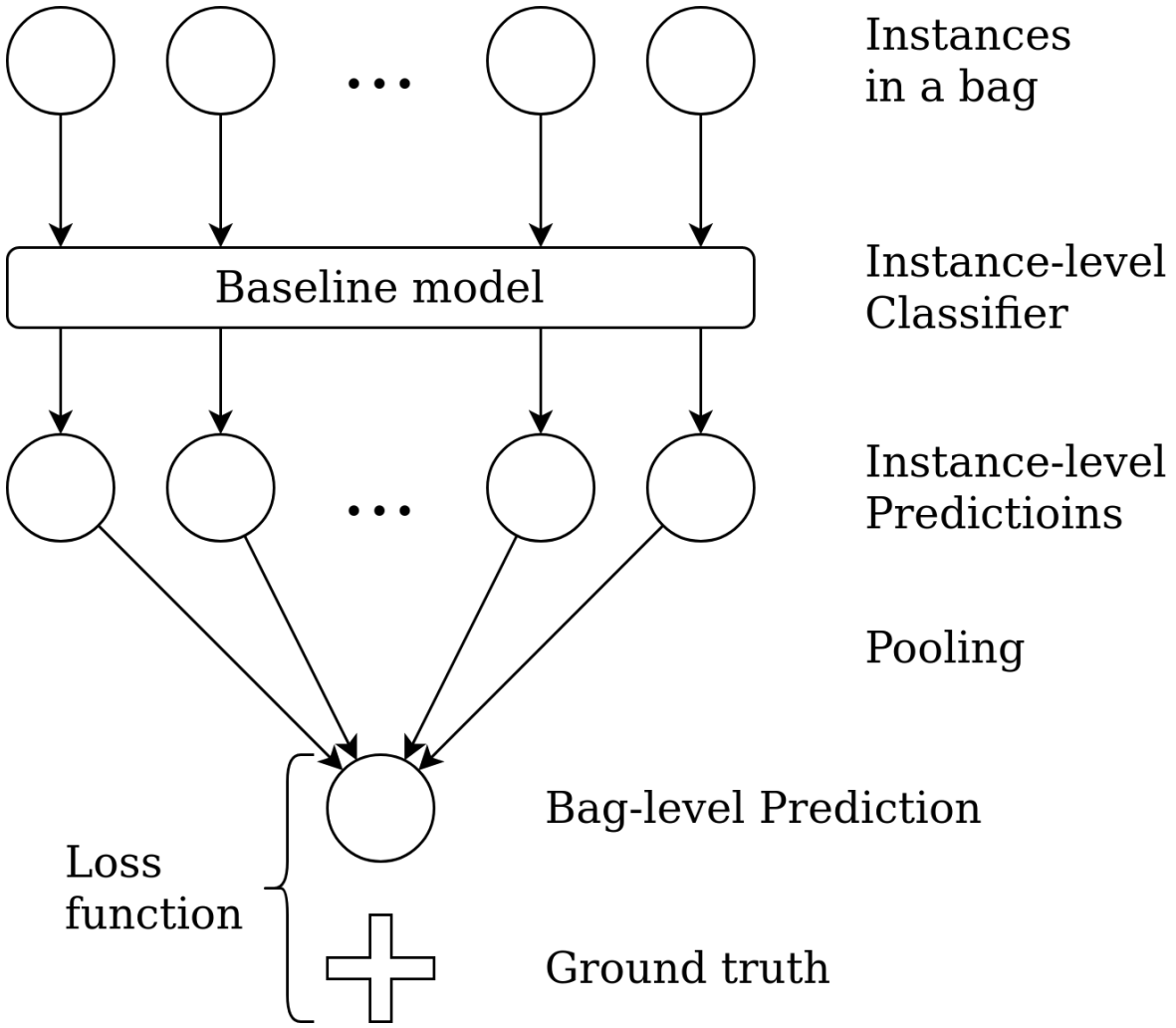


Multi Instance Learning (multi-class)

Define $0 \leq \alpha \leq 1$, denoting the proportion of elements x of X-class in a bag of *n* samples

Then Bag-label = A,
if $\frac{x}{n} \geq \alpha$

Example($\alpha = 0.75$):
[1,3,1,1,1,1,4,1] – labels of samples
Bag-label=1



Loss function

$$loss = \gamma\ CCE\left(Y_{pred}, Y_{true}\right) + (1 - \gamma)\ \mathrm{CCE}(\mathrm{Y_{pred}}, \mathrm{Y_{bag}})$$

Где $Y_{pred}, Y_{true}, Y_{bag}$ – метки классов,

γ – balance coefficient,

CCE – Categorical Cross-entropy

BERT parameters

- Model – BERT uncased_L-12_H-768_A-12
- Max_seq_length = 512 – Maximum total number of tokens (256+256)

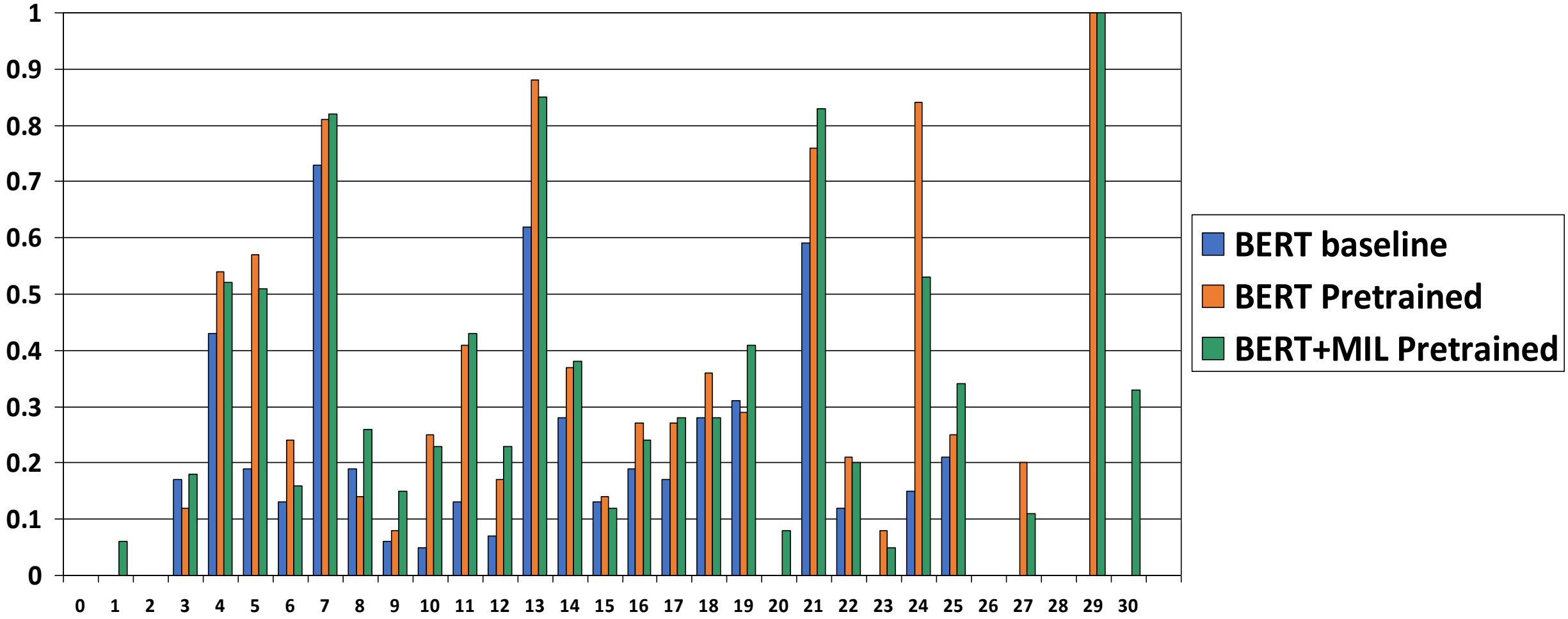
MIL:

- $\gamma \in [0.5, 0.85]$ with step=0.05 – loss functions balance coefficient
- batch_size = 6 – number of samples in a bag
- $\alpha = 0.65$ – the rate of occurrence of the target class in the bag

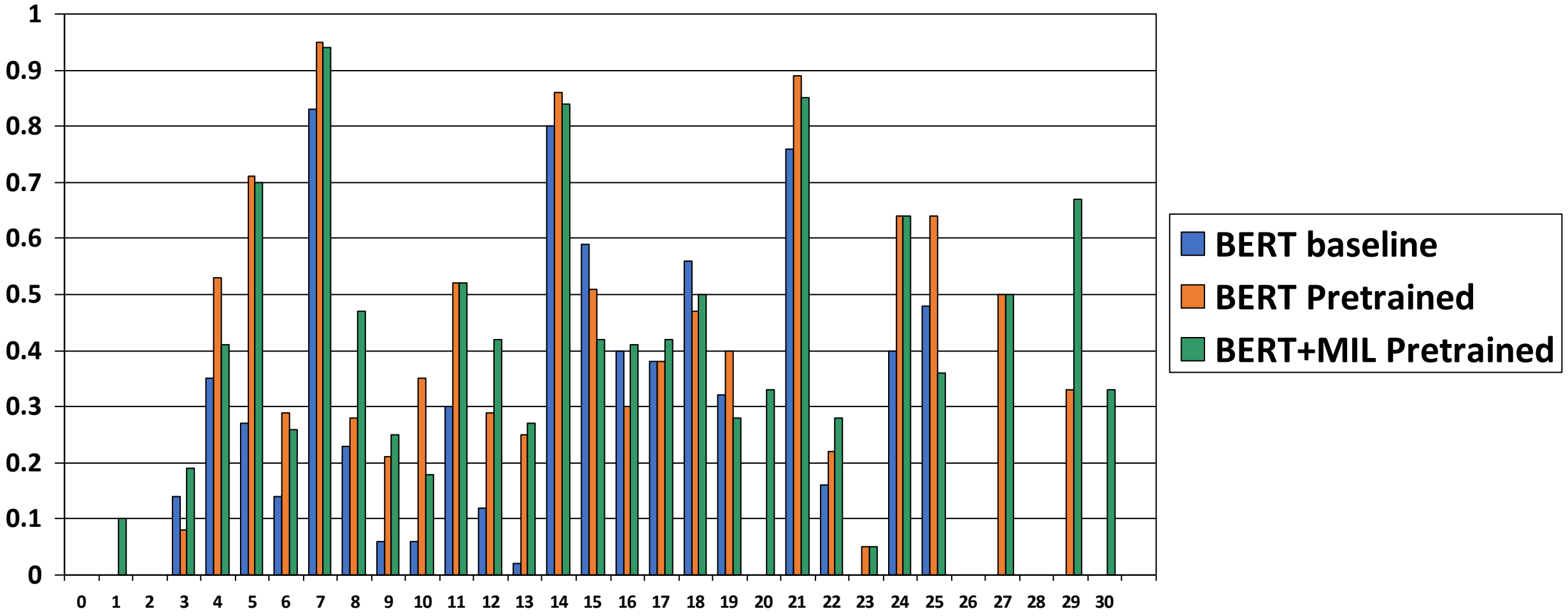
  examples: [1,1,1,1,5,27], [2,10,2,2,19,2]

- The experiments were launched in parallel at several GPUs

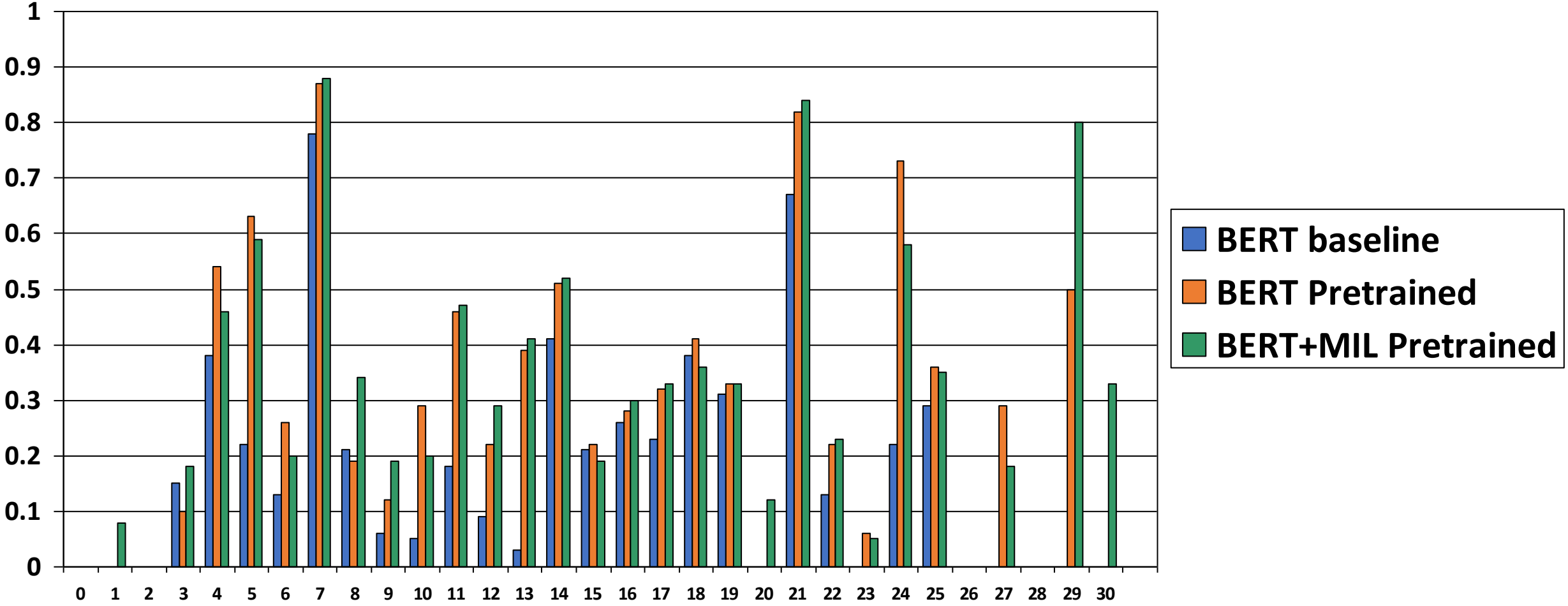
Results (precision)



Results (recall)



Results (F1-score)



Conclusion

- MIL demonstrates improvement in classification quality when learning on noisy data.
- The improvement is especially noticeable on rare classes.

Future works:

- Integration of the developed model into the text emotional charge analysis system.
- Using distributed computing to run a text analyzer.

Thanks for attention!

Sergey Volkov

volkserg1@gmail.com

Devyatkin D. A.

Shvets A. V.