Semantic Role Labeling with Neural Networks for Texts in Russian

A.O. Shelmanov, D.A. Devyatkin

shelmanov@isa.ru,

devyatkin@isa.ru

FRC CSC RAS

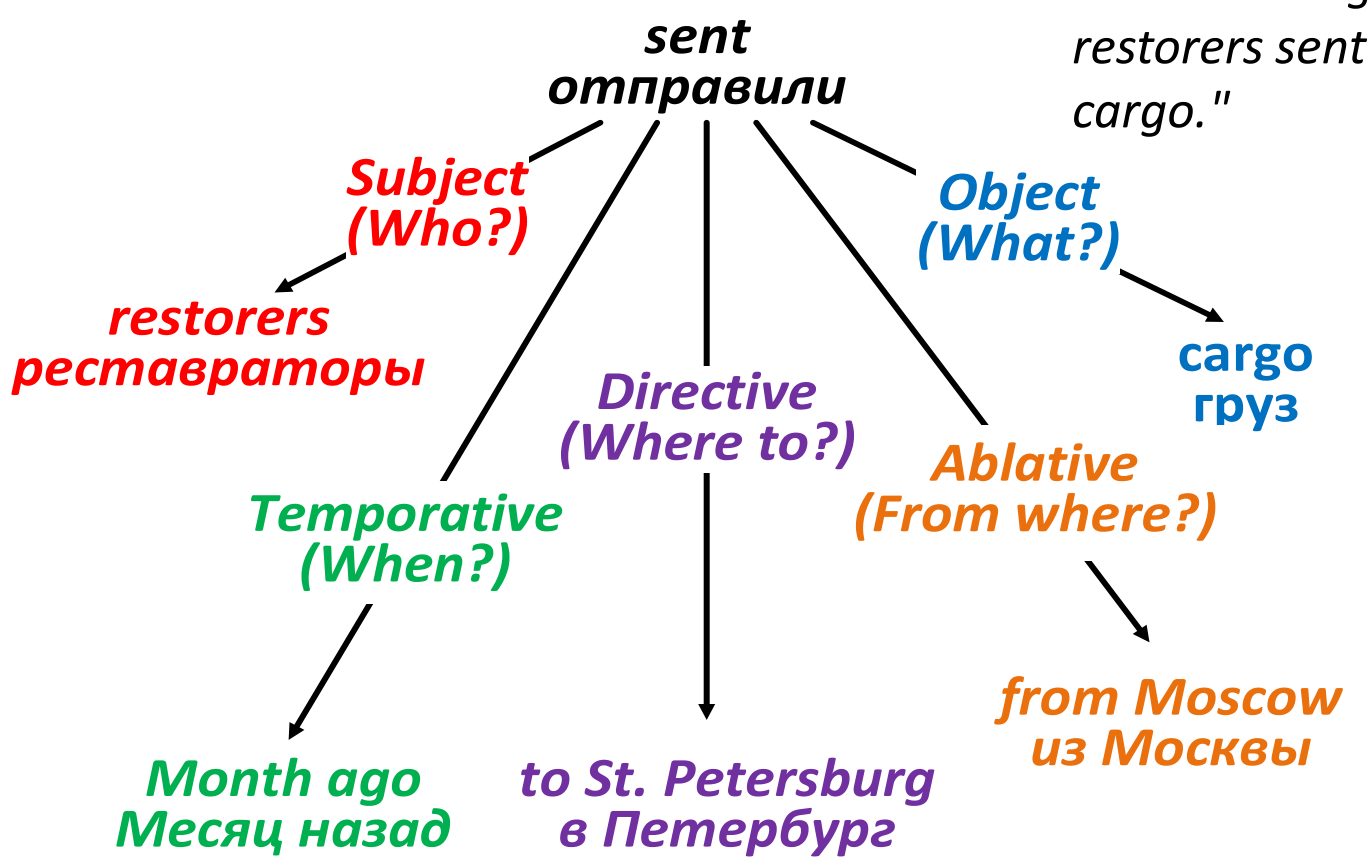
2017, Moscow

Semantic role labeling

Semantic role labeling (SRL)
(shallow semantic parsing) :
1. Determines situations in sentence
2. Identifies arguments of situations
3. Classifies arguments of situations and assigns them thematic roles

Example: *"Месяц назад из Москвы в Петербург столичные реставраторы отправили необычный груз"*

"Month ago from Moscow to St. Petersburg, metropolitan restorers sent an unusual cargo."



Applications of semantic role labeling

- Question-answering search
  - Shen D. and Lapata M., 2007
- Information extraction
  - Christensen J. et al., 2010
- Information search
  - Osipov G. et al., 2016
- Summarization
  - Khan A. et al., 2015
- Machine translation
  - Xiong D. et al., 2012
  - Bazrafshan M. and Gildea D., 2013
- Event extraction ~ Semantic role labeling
  - Wang X. et al., 2012

Related work

- Seminal work for statistical and machine learning methods for SRL:
  - Gildea D. and Jurafsky D., 2000
- Main corpora:
  - FrameNet (Baker C. F., Fillmore C. J., Lowe J. B., 1998)
  - PropBank (Kingsbury P. and Palmer M., 2002)
- Shared tasks: CoNLL 2004, 2005, 2008, 2009
  - Hajič J. et al., 2009 and other publications
- New methods based on neural networks:
  - Collobert et al., 2011
  - FitzGerald et al., 2015
  - Roth and Lapata, 2016
  - Foland W. and Martin J., 2015
  - Marcheggiani et al., 2017
  - Zhou and Xu, 2015 (end-to end)
  - Swayamdipta et al., 2016

Semantic role labeling for Russian

- Rule-based semantic parsers:
  - AOT.ru (Sokirko, A. 2001)
  - The parser of ISA FRC CSC RAS (Shelmanov and Smirnov, 2014)
  - etc.
- Known corpora annotated with semantic roles:
  - The corpus from ISA FRC CSC RAS
    - Shelmanov and Smirnov, 2014
  - FrameBank
    - Lyashevskaya, 2012
    - Lyashevskaya and Kashkin, 2015
- Data-driven semantic role labelers:
  - SVM-based parser + feature engineering (Kuznetsov I., 2015) trained on pre-release version of FrameBank
  - The parser of ISA FRC CSC RAS (Shelmanov and Smirnov, 2014) – bootstrapping based on automatic annotation of SynTagRus using rule-based semantic parser

FrameBank

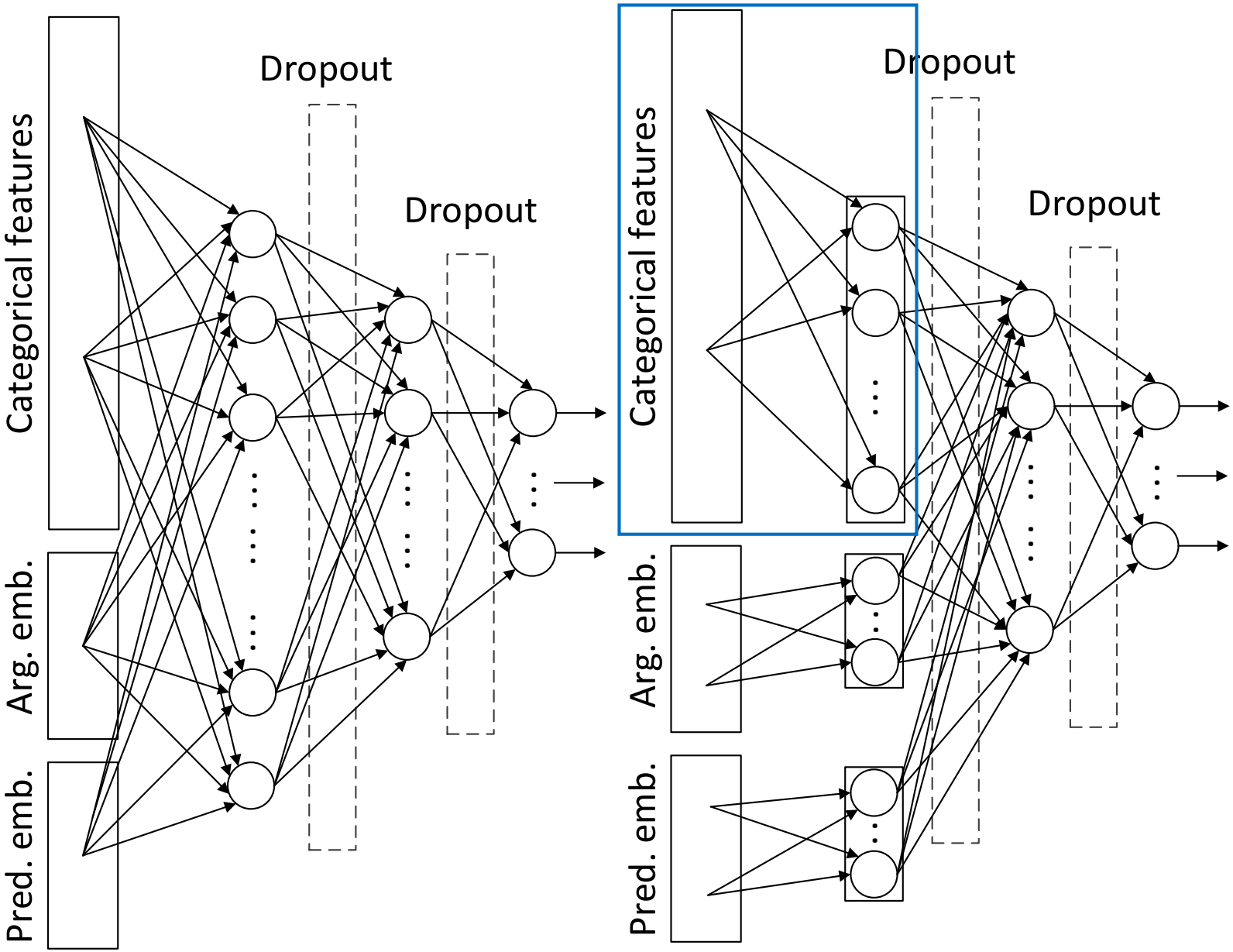
https://github.com/olesar/framebank

- Semantic role hierarchy:
  - The FrameBank provides fine and coarse grained roles
  - It also provides generality relations between roles
- Lexicon of predicate frames:
  - Describes predicates frames and their roles in terms of morphological, syntactic, semantic, and other features
- **Annotated text samples in Russian**
  - Partially annotated text samples with predicates, arguments, and semantic roles (core and non-core)
  - In addition: morphology features, lemmas, sentences, etc.
- Statistics:
  - ~ 800 unique predicates (verbs)
  - ~ 70 unique roles
  - ~ 60 000 extractable arguments

Features for semantic role labeling

- Neural networks allow to use atomic features
- Categorical features:
  - Various types of morphological features of both an argument and a predicate: part of speech, grammar case, animacy, verb form, time, passiveness, and others (“morph”)
  - Relative position of an argument in a sentence with respect to a predicate (“rel_pos”)
  - Predicate lemma (“pred_lemma”)
  - Preposition of an argument extracted from a syntax tree (“arg_prep”)
  - Name of a syntax link from an argument to its parent in a syntax tree (“synt_link”)
- Embeddings:
  - Embedding of an argument lemma (“arg_emeddings”)
  - Embedding of a predicate lemma (“pred_emeddings”)

Neural network models



A) Simple model

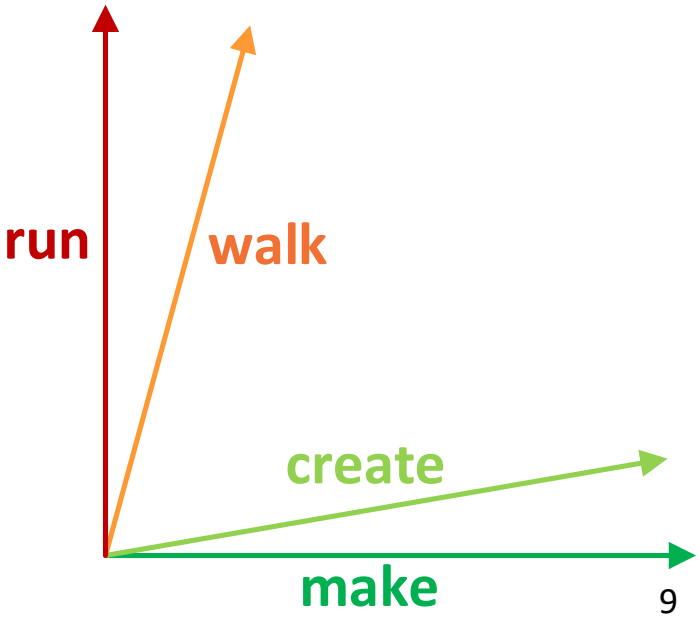
B) Complex model

Training model for “unknown” predicates

- Cannot use predicate lemma since it is not available in the training corpus for “unknown” predicates
- Instead of predicate lemma we are using word embeddings built using word2vec
  - Embeddings encode semantic similarities of words in a low dimensional vector space

⇩

  - Embeddings can encode similarities between predicate frames
- We used RusVectores 2.0 models (Kutuzov and Andreev, 2015):
  - Trained on Russian national corpus
  - 300 dimensions
- Used early stopping during training of neural networks

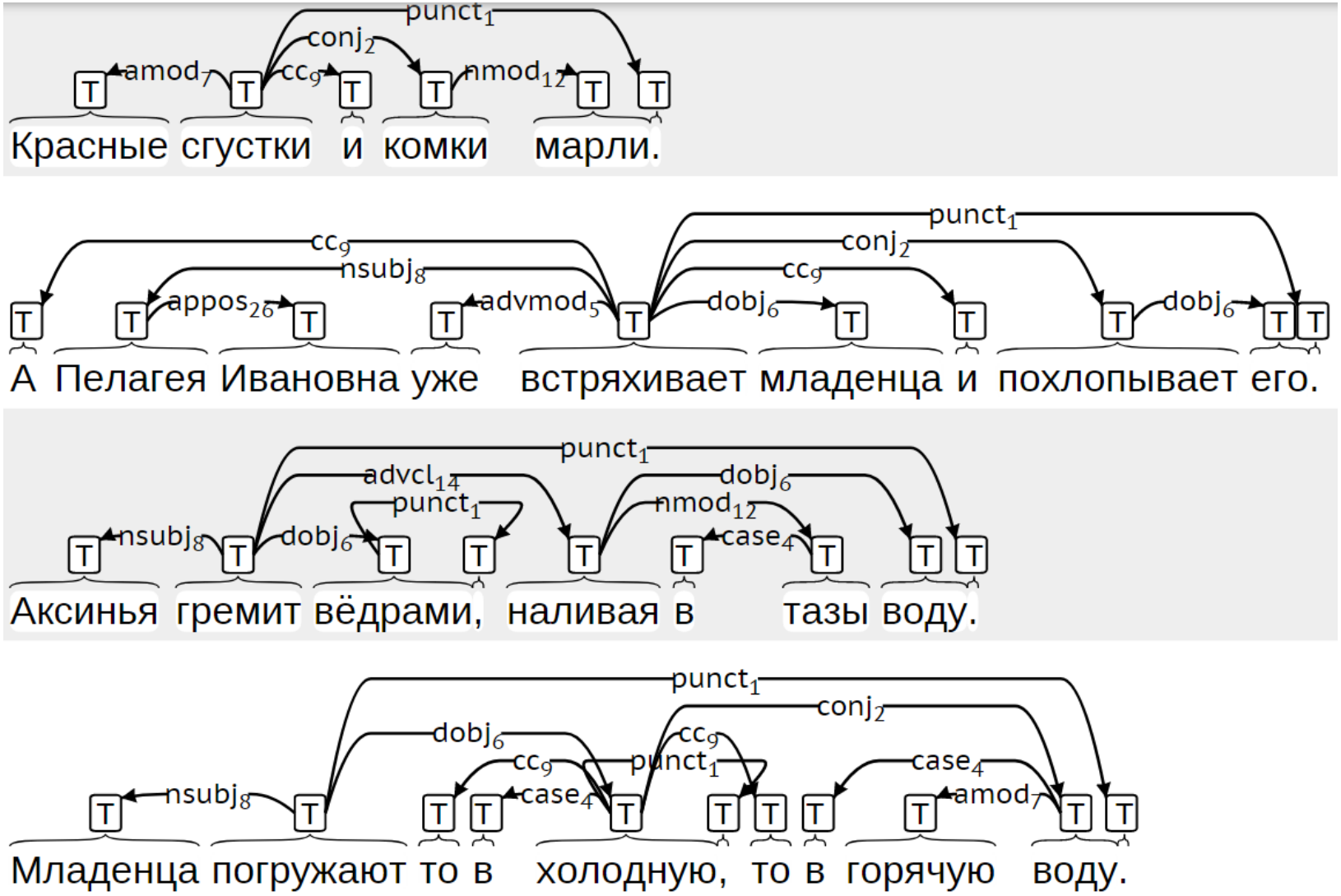


Data preparation

- Original FrameBank does not provide explicit correspondence between text offsets and SRL annotations
  - We created the automatic tool for mapping predicates and arguments with core roles to text tokens http://nlp.isa.ru/framebank_parser
- Parsed with Google's SyntaxNet
  - For parsing we used SyntaxNet (McParsey model for Russian) (Andor D., 2016)
  - We prepared dockerized version of SyntaxNet for Russian (and other languages) + Python wrapper
    - https://github.com/IINemo/docker-syntaxnet_rus
      bash$> echo "мама мыла раму" | docker run --rm -i inemo/syntaxnet_rus
    - https://github.com/IINemo/syntaxnet_wrapper
  - The syntax structure corresponds to well-known Universal dependencies format

Example of FrameBank parsing via SyntaxNet

- http://nlp.isa.ru/brat_framebank/index.xhtml#/framebank_part_syntaxnet/



Experiment setup for training of a model for "known predicates"

- Selected the subcorpus by keeping only predicates with > 10 examples
  - 572 predicates left
- Filtered infrequent semantic roles and fixed erroneous role labels
- Final version of experimental dataset:
  - 53,151 examples
  - 44 different semantic roles
- RusVectores problem:
  - Large portion of predicates (verbs) are not covered
  - 17,000 examples in our dataset have zero predicate embeddings
- Baseline: most frequent role in the corpus
- Five-fold cross-validation

Results of neural-network models for "known" predicates

Model + feature set	Macro F_1-score,%	Micro F_1-score,%
Baseline	0.5 ± 0.0	11.6 ± 0.2
Simple + morph	22.8 ± 0.6	35.4 ± 0.3
Simple + morph + pred_lemma	71.2 ± 0.6	76.1 ± 0.5
Simple + morph + pred_emeddings	62.0 ± 0.4	65.2 ± 0.3
Simple + morph + pred_lemma + arg_prep	75.9 ± 0.4	79.2 ± 0.2
Simple + morph + pred_lemma + arg_prep + synt_link	76.8 ± 0.5	80.3 ± 0.3
Simple + morph + pred_lemma + arg_prep + synt_link + arg_embeddings + pred_embeddings	78.6 ± 0.4	81.8 ± 0.2
Complex + morph + synt + pred_lemma + arg_embeddings + pred_embeddings	**79.2 ± 0.3**	**82.3 ± 0.2**

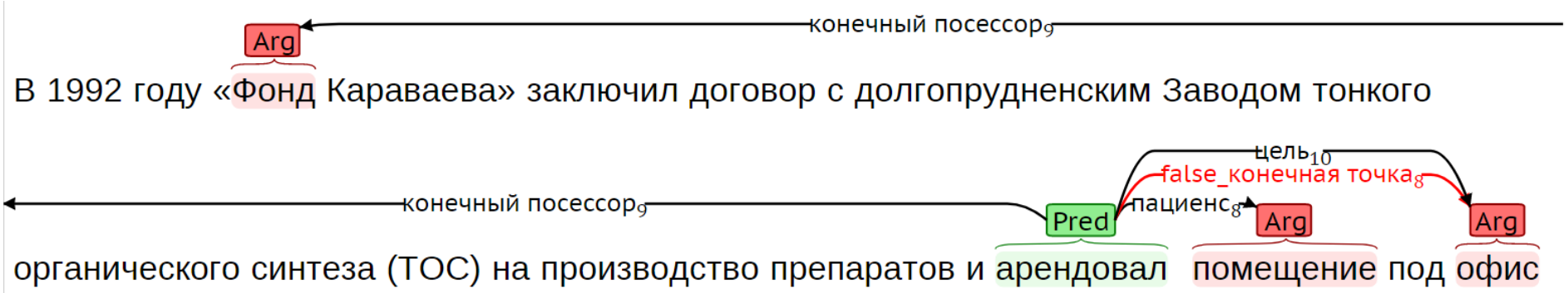
Results of non-neural-network models for "known" predicates

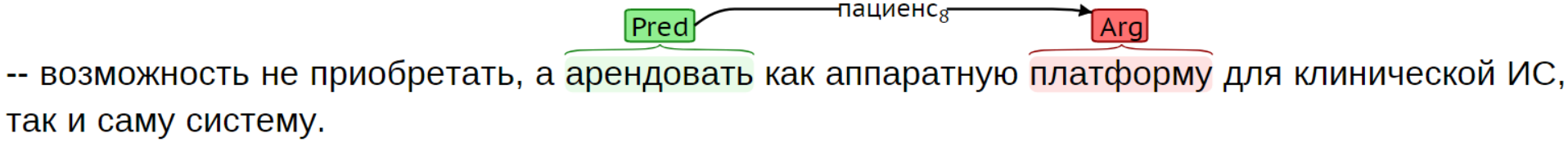
- Used the most complete feature set
- Tuned parameters using greedy strategy

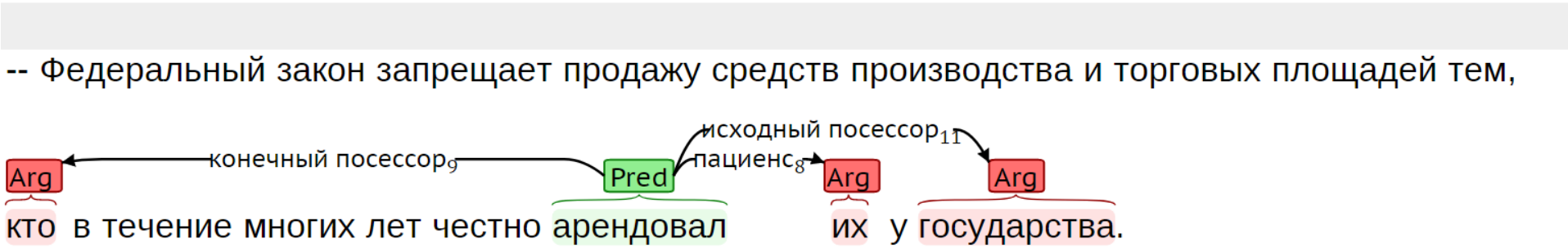
Model	Macro F_1-score,%	Micro F_1-score,%
LinearSVC	74.3 ± 0.2	77.6 ± 0.1
LogReg	75.1 ± 0.1	78.2 ± 0.3
LightGBM	71.3 ± 0.4	76.0 ± 0.1
Random Forest	69.7 ± 0.4	71.9 ± 0.1
+Top neural network	**79.2 ± 0.3**	**82.3 ± 0.2**

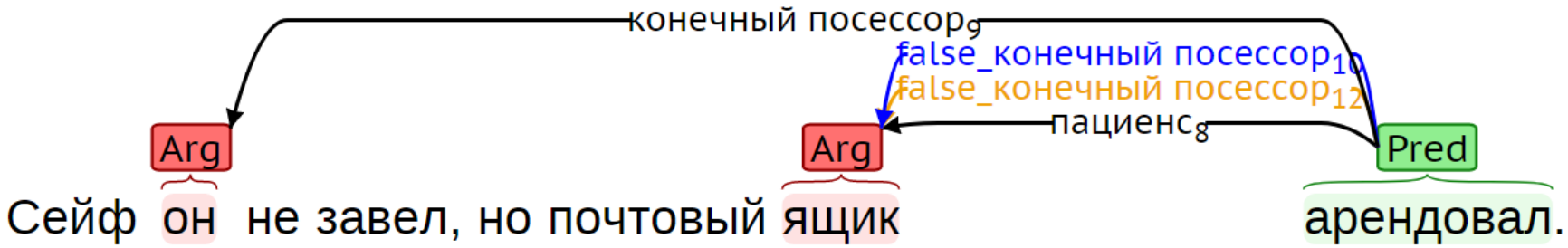
Example of SRL parsing

- http://nlp.isa.ru/brat_framebank









Experiments for “unknown” predicates

- The sets of **predicates** for training and testing **do not intersect**
- Performed evaluations for two different split methods:
  - **The good split:**
    - The test set contains highly similar predicates to the ones in the training set (by cosine similarity)
    - Easy for the models to restore semantic frame for “unknown” predicate
    - Training set: 49,709 examples
    - Testing set: 27 predicates; 3,442 examples
  - **The bad split**:
    - Test set contains predicates that are least similar to any of the “known” predicates
    - Hard for the models to restore semantic frame for “unknown” predicate
    - Training set: 50,093 examples
    - Testing set: 21 predicates; 3,058 examples
- Averaged across several fits using different random states

Results of models for “unknown” predicates

- **Results for the “good” split**

Model + feature set	Macro F_1-score,%	Micro F_1-score,%
Baseline	0.4	9.6
Simple (only categ. Feats)	13.7 ± 0.4	24.6 ± 0.3
Complex + arg_embeddings	19.4 ± 0.3	31.9 ± 0.5
Complex + arg_pred_embeddings	41.4 ± 0.7	66.7 ± 1.1

- **Results for the “bad” split**

Model + feature set	Macro F_1-score,%	Micro F_1-score,%
Baseline	0.7	13.2
Simple (only categ. Feats)	9.1 ± 0.2	24.8 ± 0.5
Complex + arg_embeddings	14.5 ± 0.7	27.2 ± 0.1
Complex + arg_pred_embeddings	24.1 ± 1.5	41.4 ± 2.2

Created electronical resources

- Preprocessed FrameBank + scripts + models:
  - http://nlp.isa.ru/framebank_parser
- Original FrameBank + results of SRL and syntax parsing visualized via Brat tool (Stenetorp P. et al., 2012):
  - http://nlp.isa.ru/brat_framebank
- Dockerized version of SyntaxNet for Russian:
  - **bash$>** echo "мама мыла раму" | docker run --rm -i inemo/syntaxnet_rus
  - https://github.com/IINemo/docker-syntaxnet_rus
- Python wrapper for SyntaxNet:
  - https://github.com/IINemo/syntaxnet_wrapper
- Dockerized version of SyntaxNet for English:
  - **bash$>** echo "Beware of a silent dog and still water" | docker run --rm -i inemo/syntaxnet_eng
- Docker containers with SyntaxNet for other languages:
  - https://github.com/IINemo/docker-syntaxnet

Conclusion and future work

- **Results:**
  - Presented the neural network models for semantic role labeling of Russian texts
  - Investigated the method for training a labeler for arguments of “unknown” predicates using word embeddings
  - Created benchmark based on FrameBank corpus for evaluation of parsers for SRL
  - The models and the benchmark are openly available
- **Future work:**
  - Adding global inference step for the SRL parser
  - Methods of merging several semantic parsers and their annotations for creating a corpus with a higher annotation recall
  - New methods base on semi-supervised learning
- **Acknowledgments**
  - The project is supported by the Russian Foundation for Basic Research, project number: 16-37-00425 “mol_a”.

References (1)

- Andor D., Alberti C., Weiss D., Severyn A., Presta A., Ganchev K., Petrov S., and Collins M. (2016), Globally normalized transition-based neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2442–2452.
- Baker C. F., Fillmore C. J., Lowe J. B. The Berkeley FrameNet project //Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – 1998. – P. 86-90.
- Bazrafshan M., Gildea D. Semantic Roles for String to Tree Machine Translation // ACL. – 2013. – P. 419-423.
- Christensen J. et al. Semantic role labeling for open information extraction //Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading. –2010. – P. 52-60.
- Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., and Kuksa P. (2011), Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug):2493–2537.
- Gildea D., Jurafsky D. Automatic labeling of semantic roles //Computational linguistics. – 2002. – P. 245-288.
- Hajič J. et al. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages //Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task. – 2009. – P. 1-18.

References (2)

- Khan A., Salim N., Kumar Y. J. A framework for multi-document abstractive summarization based on semantic role labelling //Applied Soft Computing. – 2015. – P. 737-747.
- Kingsbury P., Palmer M. From TreeBank to PropBank //LREC. – 2002. – P. 1989-1993.
- Kuznetsov I. (2015), Semantic role labeling for Russian language based on Russian FrameBank. In Proceedings of International Conference on Analysis of Images, Social Networks and Texts, P. 333–338.
- Lyashevskaya O. and Kashkin E. (2015), FrameBank: a database of Russian lexical constructions. In International Conference on Analysis of Images, Social Networks and Texts, pp. 350–360.
- Lyashevskaya O. (2012), Dictionary of valencies meets corpus annotation: a case of Russian FrameBank. In Proceedings of the 15th EURALEX International Congress, volume 15.
- Osipov G. et al. Exactus Expert—Search and Analytical Engine for Research and Development Support //Novel Applications of Intelligent Systems. – 2016. – P. 269-285.

References (3)

- Shen D., Lapata M. Using Semantic Roles to Improve Question Answering //EMNLP-CoNLL. – 2007. – P. 12-21.
- Sokirko, A. (2001), A short description of Dialing Project, available at: http://www.aot.ru/
- Stenetorp P. et al. BRAT: a web-based tool for NLP-assisted text annotation //Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. – 2012. – P. 102-107.
- Wang X., Gerber M. S., Brown D. E. Automatic crime prediction using events extracted from twitter posts //International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. –2012. – p. 231-238.
- Xiong D., Zhang M., Li H. Modeling the translation of predicate-argument structure for SMT //Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. – 2012. – P. 902-911.