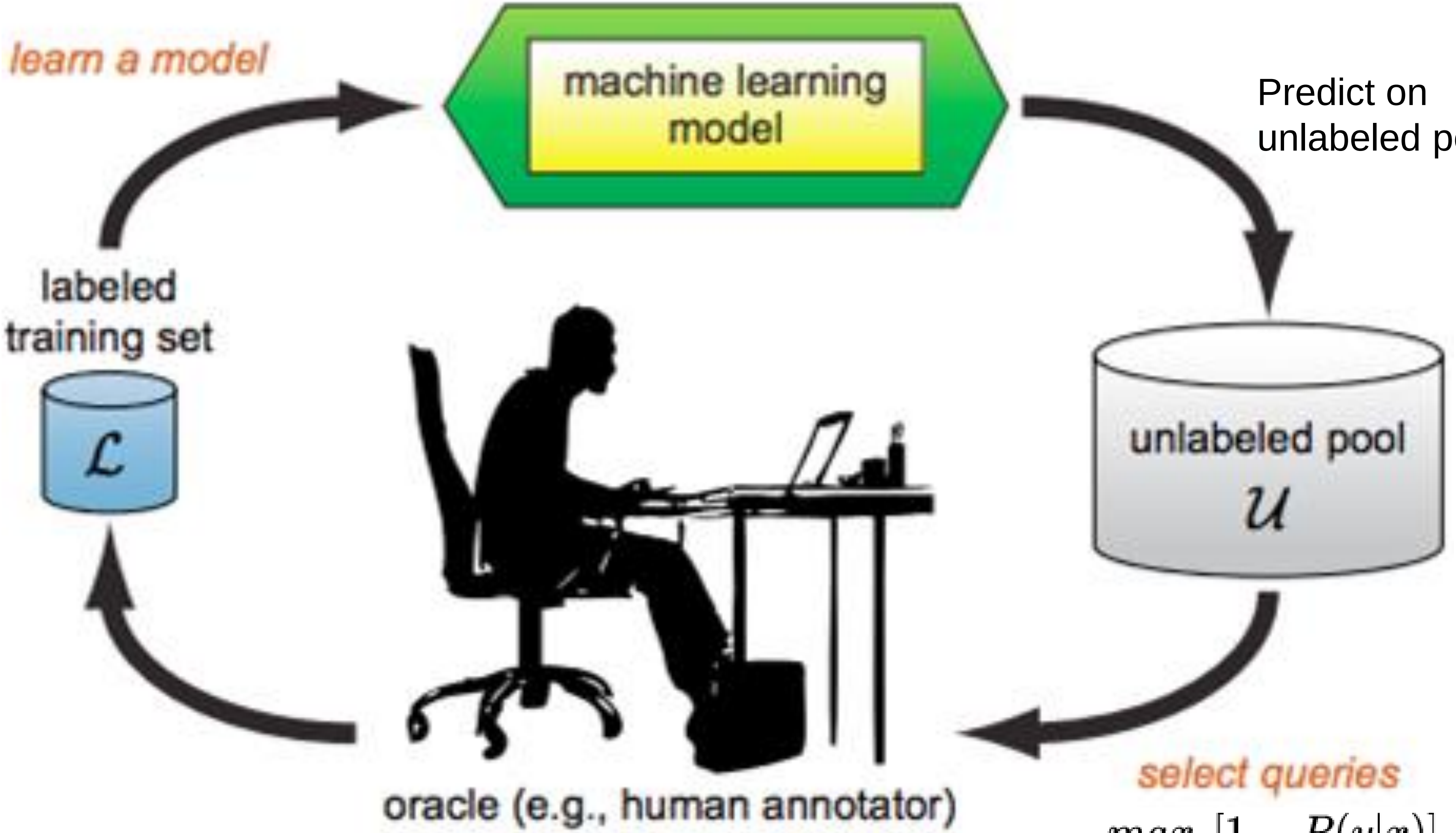
Deep Active Learning: Reducing Annotation Effort for Automatic Sequence Tagging of Clinical and Biomedical Texts

Dr. Artem Shelmanov
Research Scientist @ Skoltech

Skoltech
Skolkovo Institute of Science and Technology

Basic Idea of Active Learning (AL)



Predict on unlabeled pool

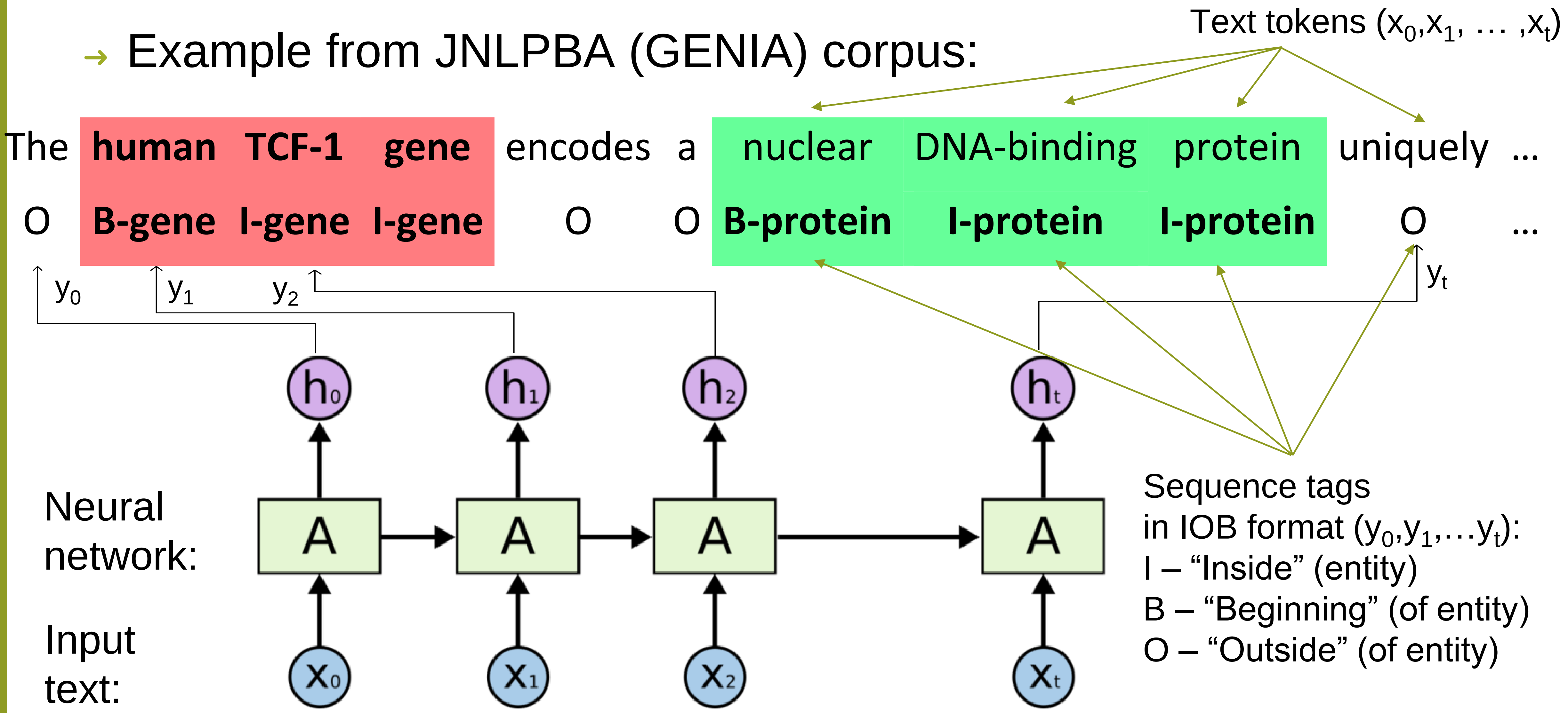
$$max_y[1 - P(y|x)]$$

(from Burr Settles et al.)

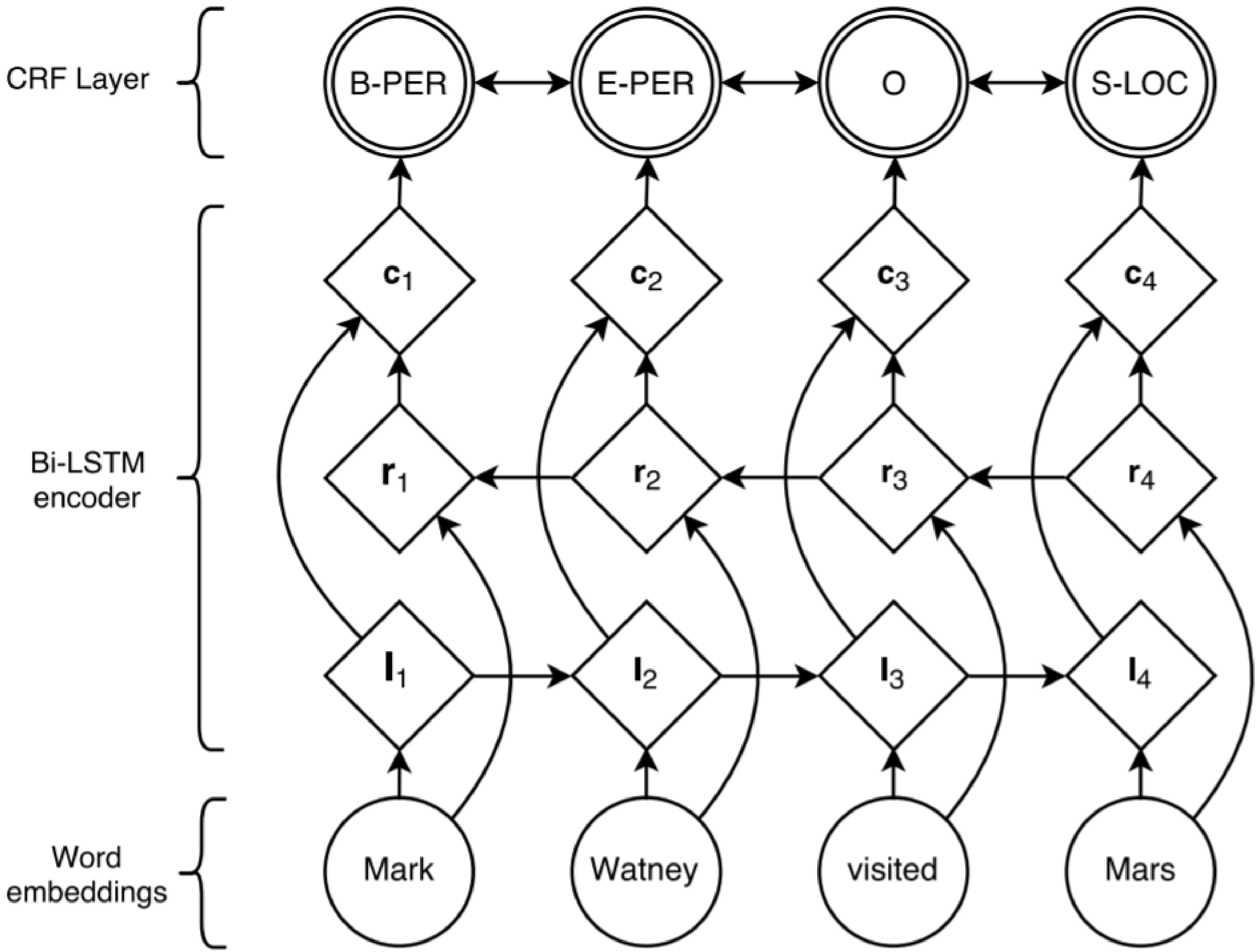
Skoltech
Skolkovo Institute of Science and Technology

Sequence Tagging Task (NER)

→ Example from JNLPBA (GENIA) corpus:



Popular Architecture



- BiLSTM-CRF (Ma and Hovy, 2016)
- Near SOTA results if accompanied with strong word representations

Classical AL
Query Strategies

Common Query Strategies: Uncertainty Sampling (Lewis and Catlett, 1994)

→ Uncertainty sampling: the learner queries the instance, about which it has the least certainty

Least confidence (McCallum et al., 2005):

$$\phi^{LC}(\mathbf{x}) = 1 - P(\mathbf{y}^*|\mathbf{x};\theta)$$

Margin (Scheffer et al., 2001):

$$\phi^{M}(\mathbf{x}) = -(P(\mathbf{y}_1^*|\mathbf{x};\theta) - P(\mathbf{y}_2^*|\mathbf{x};\theta))$$

Token entropy:

$$\phi^{TE}(\mathbf{x}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} P_\theta(y_t = m)\log P_\theta(y_t = m)$$

N-best sequence entropy (NSE):
(Kim et al., 2006)

$$\phi^{NSE}(\mathbf{x}) = -\sum_{\hat{\mathbf{y}}\in\mathcal{N}} P(\hat{\mathbf{y}}|\mathbf{x};\theta)\log P(\hat{\mathbf{y}}|\mathbf{x};\theta)$$

Common Query Strategies: Query by Committee (Seung et al., 1992)

→ Query-by-committee: a “committee” of models selects the instance about which its members most disagree

Vote entropy (Dagan and Engelson, 1995):

$$\phi^{VE}(\mathbf{x}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\frac{V(y_t, m)}{C}\log\frac{V(y_t, m)}{C}$$

$V(y_t, m)$ – number of votes for position t and label m

Largest KL-divergence between committee members and consensus (McCallum and Nigam, 1998):

$$\phi^{KL}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C} D\left(\theta^{(c)} \| \mathcal{C}\right)$$

Sequence vote entropy:

$$\phi^{SVE}(\mathbf{x}) = -\sum_{\hat{\mathbf{y}} \in \mathcal{N}^c} P(\hat{\mathbf{y}}|\mathbf{x}; \mathcal{C}) \log P(\hat{\mathbf{y}}|\mathbf{x}; \mathcal{C})$$

Fraction of models that disagree with the most popular choice (Shen et al., 2018):

$$f_i = 1 - \frac{\max_y \left|\left\{m : \operatorname{argmax}_{y'} \mathbb{P}^m[y_i = y'] = y\right\}\right|}{M}$$

See (Settles and Craven, 2008) for further detail

Skoltech
Skolkovo Institute of Science and Technology

Problems with QbC and US Methods

- Query-by-committee is slow since you need to train an ensemble of classifiers and perform inference on all of them
- Uncertainty estimates via standard US methods are not very good for unseen regions
- Both US and QbC prone to sample outliers – objects that are useless for training a model

Several SOTA Approaches in DAL for Information Extraction

Shen et al., 2018 (ICLR-2018) (1)

“Deep active learning for named entity recognition” (Shen et al., 2018)

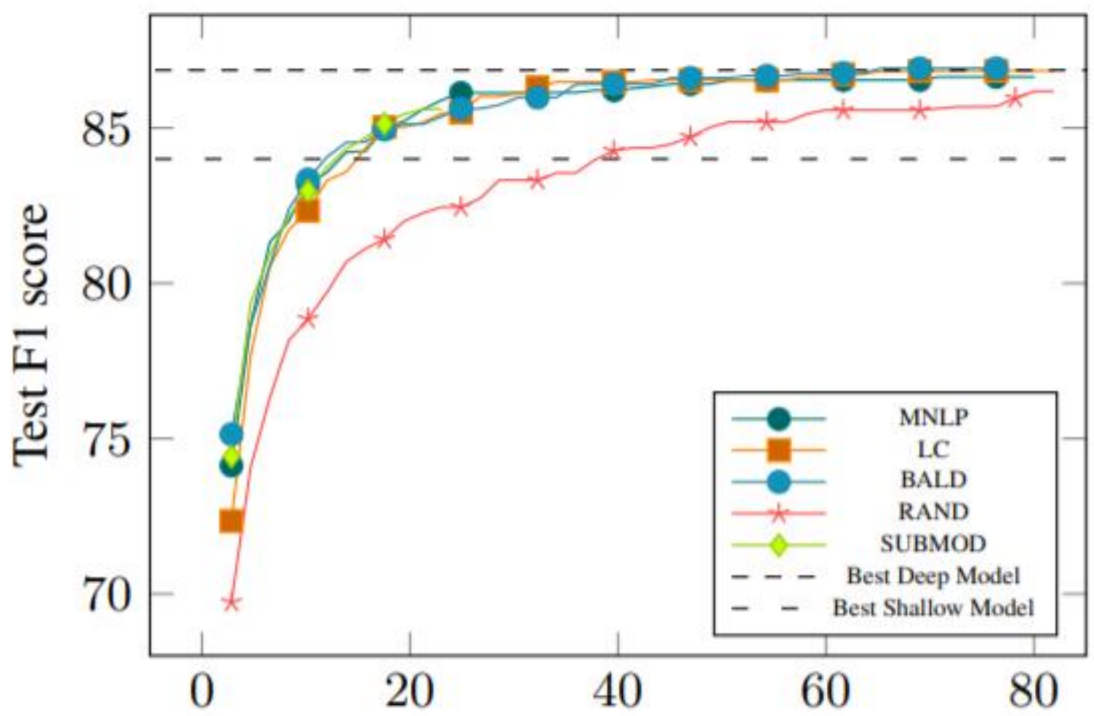
- → First work that uses deep learning model for sequence labeling in conjunction with active learning
- → Propose US strategy Maximum Normalized Log-Probability (MNLP):

$$\phi^{\mathrm{MNLP}}(x) = \max_{\{y_j\}} \frac{1}{n} \sum_{i}^{n} \log P(y_i | \{y_j\} \setminus y_i, \{x_j\})$$

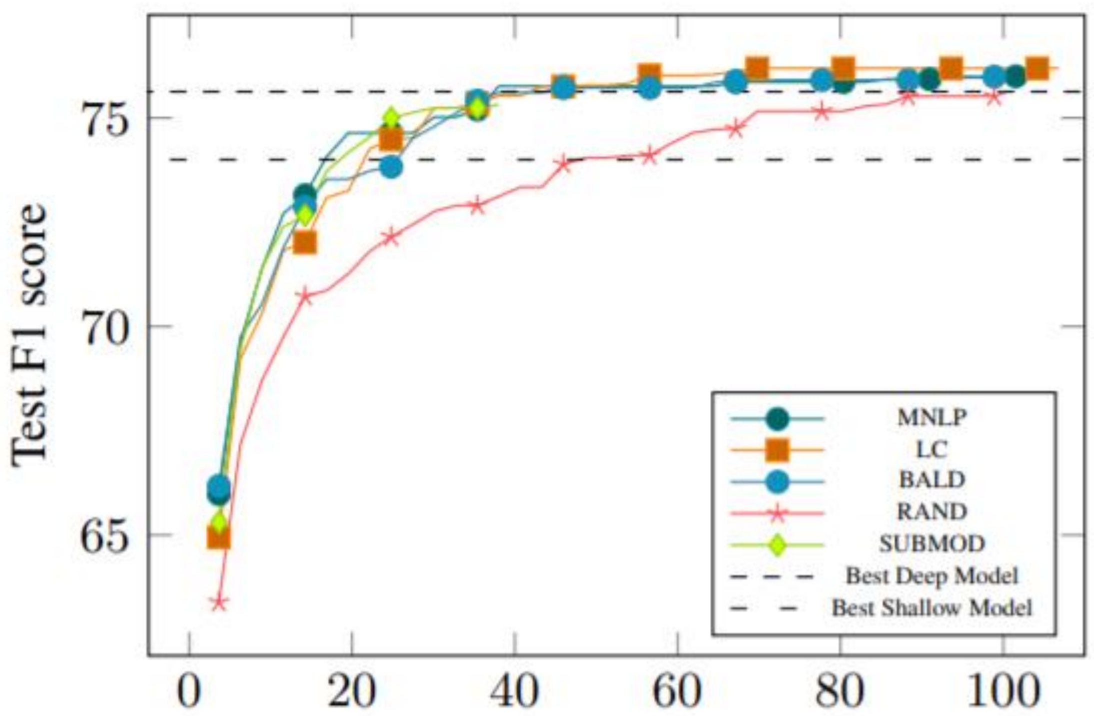
- → Propose CNN-CNN-LSTM architecture (CNN character encoder, CNN token encoder, LSTM decoder), argue that it is faster than alternatives like LSTM-LSTM-CRF

Skoltech
Skolkovo Institute of Science and Technology

Shen et al., 2018 (ICLR-2018) (2)



(a) OntoNotes-5.0 English

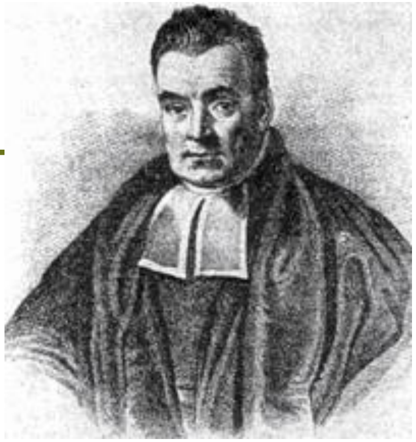


(b) OntoNotes-5.0 Chinese

- → Deep models outperform shallow
- → AL **achieves 99%** performance of the best deep model trained on full data **using only 24.9%** of data on the English dataset and 30.1% on Chinese dataset

Skoltech
Skolkovo Institute of Science and Technology

Siddhant and Lipton, 2018 (EMNLP-2018) (1)



“Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study” (Siddhant and Lipton, 2018)

→ Monte Carlo dropout (Gal et al., 2017)

- We can make several varying predictions using dropout on inference
- Quality of estimates:

“least confident” < **“Monet Carlo dropout QbC”** < “QbC on ensemble”

→ Deep Bayesian active learning (Bayes by backprop)

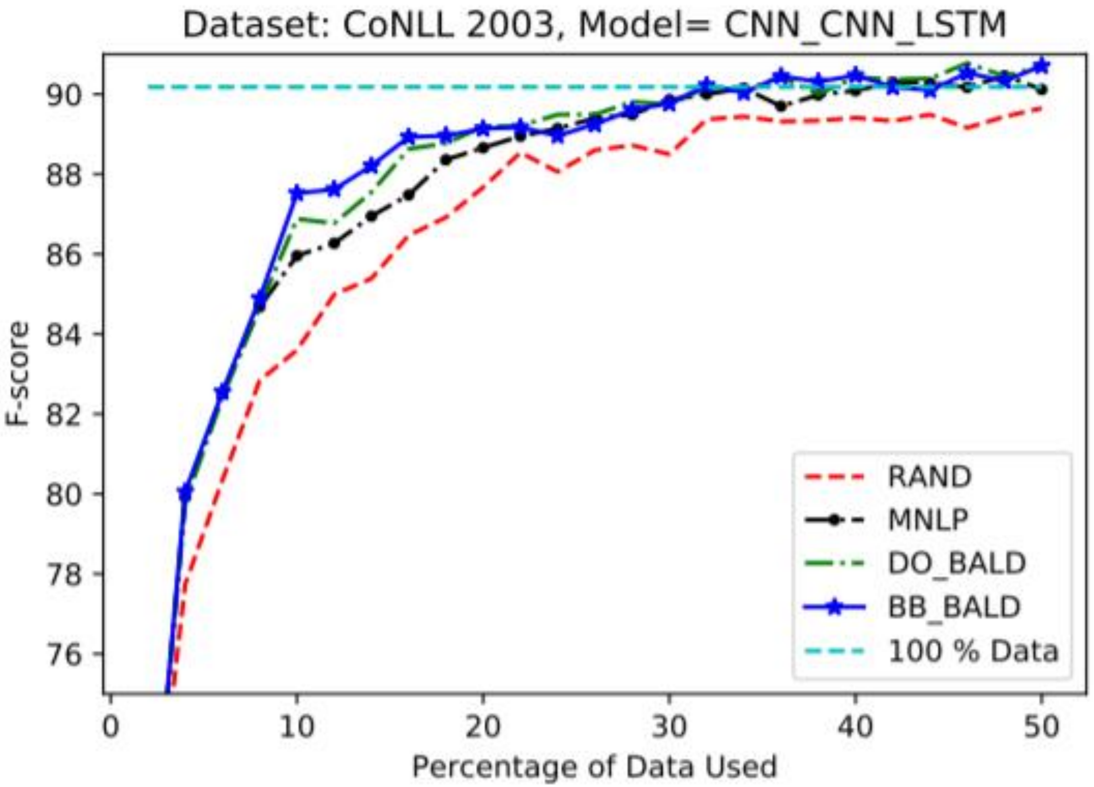
- Use Bayesian NN that maintains a probability distribution over model parameters
- Perform variational inference to obtain posterior, use MC to get uncertainty estimates

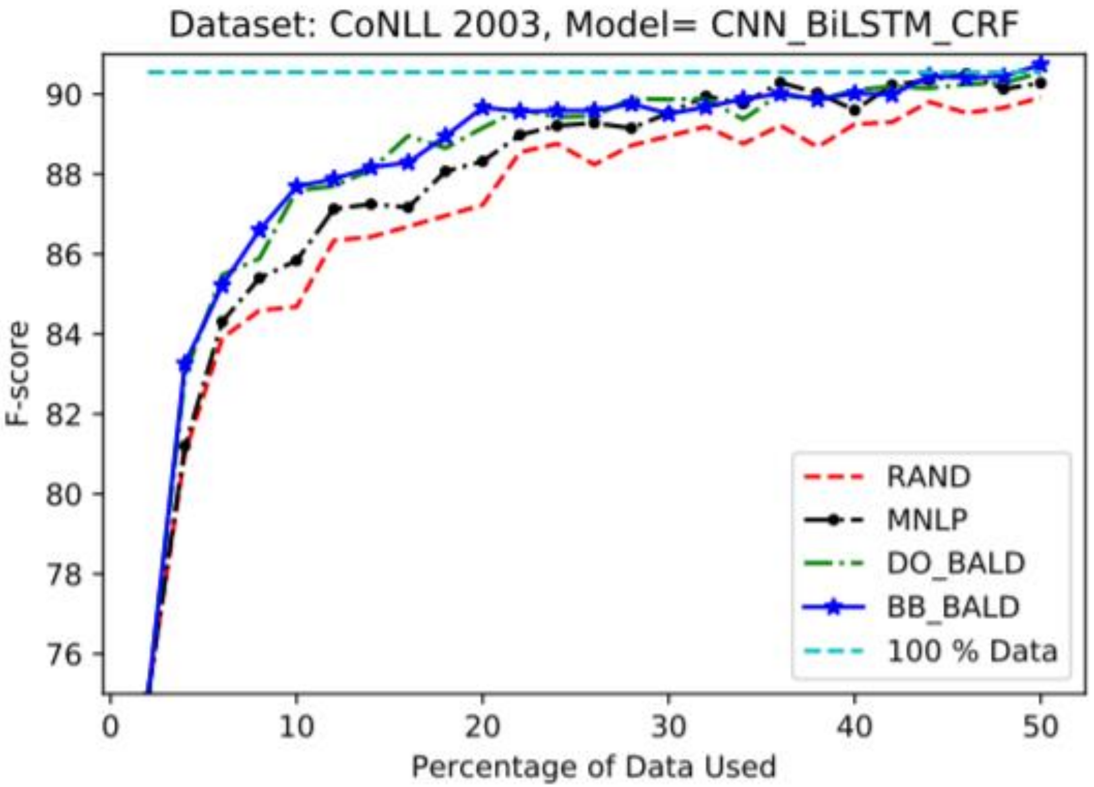
Siddhant and Lipton, 2018 (EMNLP-2018) (2)

- → Bayesian AL by disagreement (BALD):

$$f_i = 1 - \frac{\max_y \left|\{m : \mathrm{argmax}_{y'} \mathbb{P}^m[y_i = y'] = y\}\right|}{M}$$

- → Architectures: CNN-CNN-LSTM, CNN-BiLSTM-CRF
- → Experiments on CoNLL-2003, OntoNotes 5.0, and datasets for SRL and sentence classification





Bayesian > Least Confidence

Erdmann et al., 2019 (NAACL-2019)

<u>Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities (Erdmann et al., 2019)</u>

→ Novel Pre-Tag DeLex algorithm

- → Gazetteers to bootstrap annotation and to detect novel objects
- → 3 delexicalized models trained on subsets manually labeled data and automatically labeled data. => Bootstrapping cycle:

1. Use extracted objects to label data and <u>detect novel contexts</u> for objects
2. Learn contexts and use them to <u>detect novel objects</u>
3. Use extracted objects to label data and <u>detect novel contexts</u> for objects
4. …

→ Compared to: MNLP

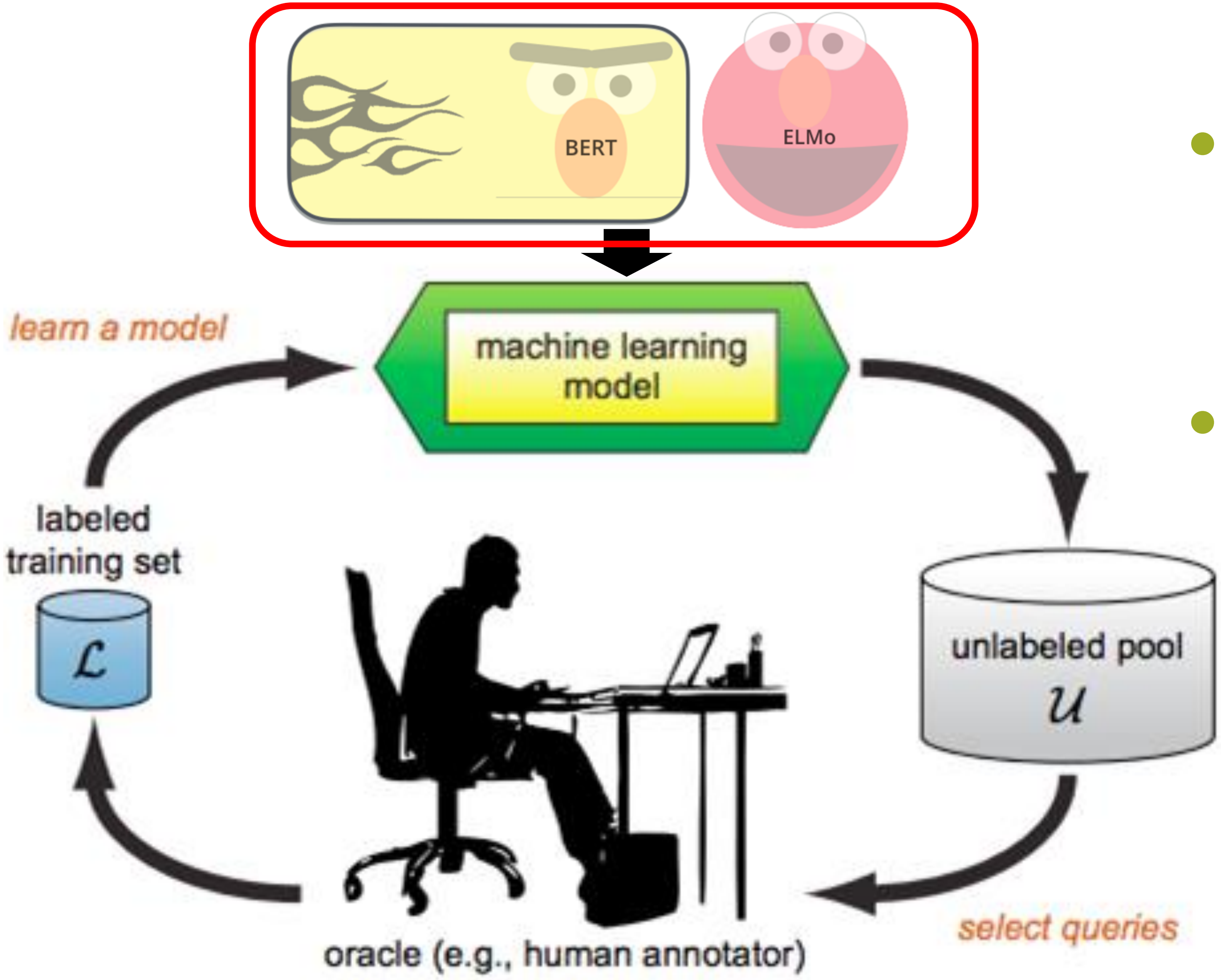
→ Architectures: BiLSTM-CRF, CNN-BiLSTM, and pure CRF

→ Experiments on Spanish CoNLL, GermEval, Arabic and Latin corpora

Active Learning with Deep Pre-trained Models for Sequence Tagging of Clinical and Biomedical Texts (IEEE BIBM 2019)

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Dmitry Dylov

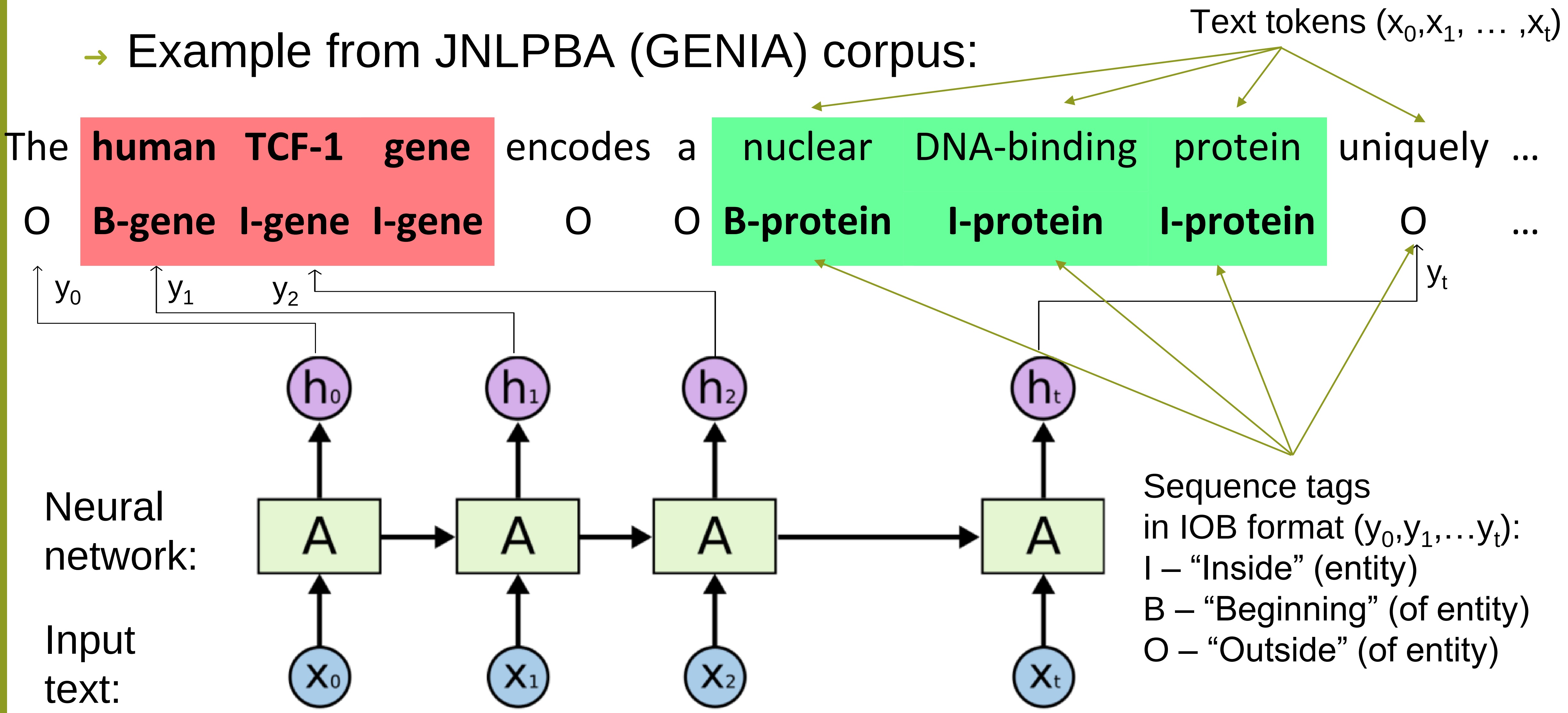
Basic Idea



- AL for IE with transfer learning:
  - → Deep pre-trained models BERT, ELMo, etc.
- Transfer learning:
  - → Provides universal feature set
  - → Enables neural network training on small datasets
  - → Very powerful for streamline NLP tasks

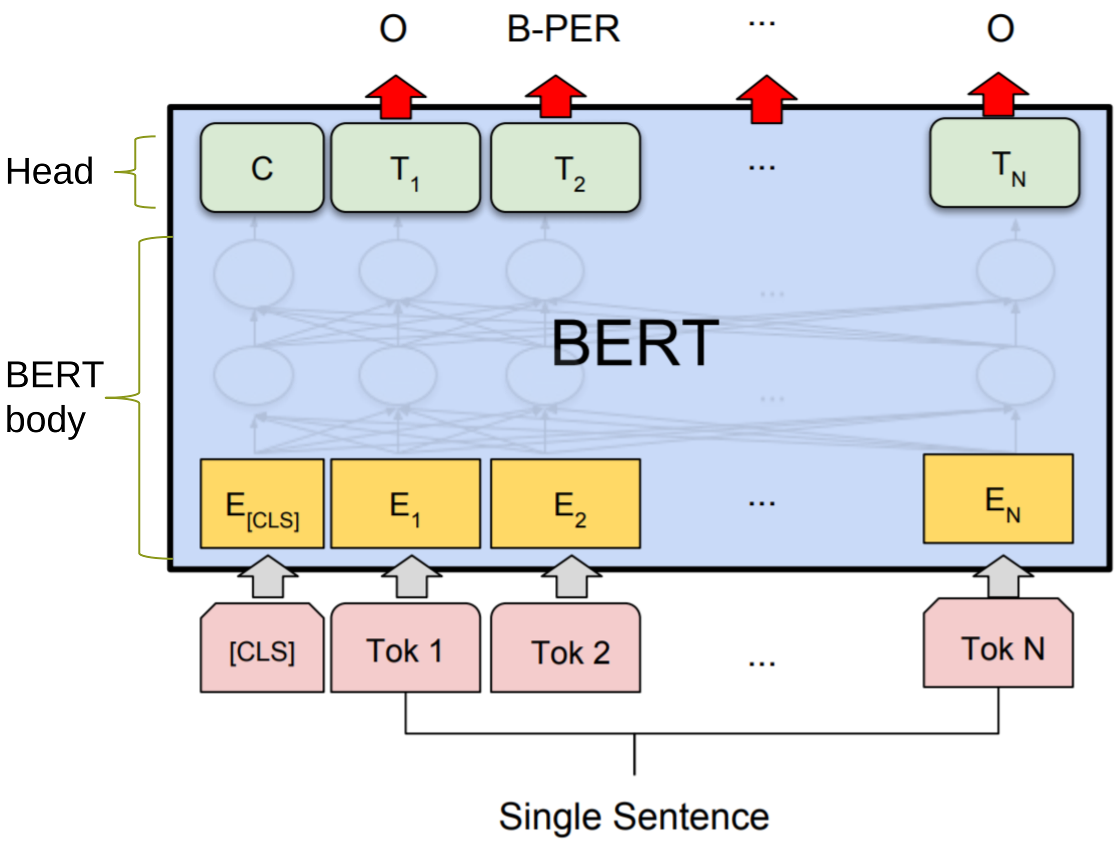
Sequence Tagging Task (NER)

→ Example from JNLPBA (GENIA) corpus:



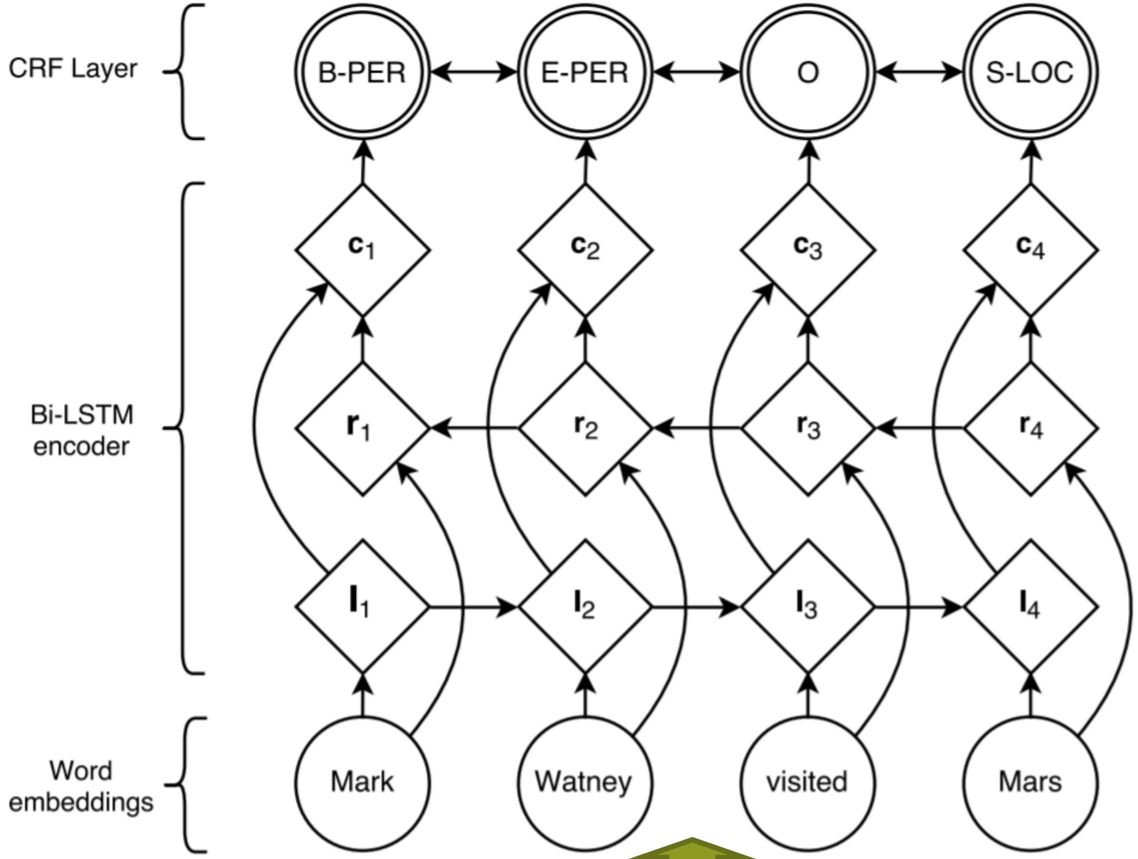
Skoltech
Skolkovo Institute of Science and Technology

NN Architectures

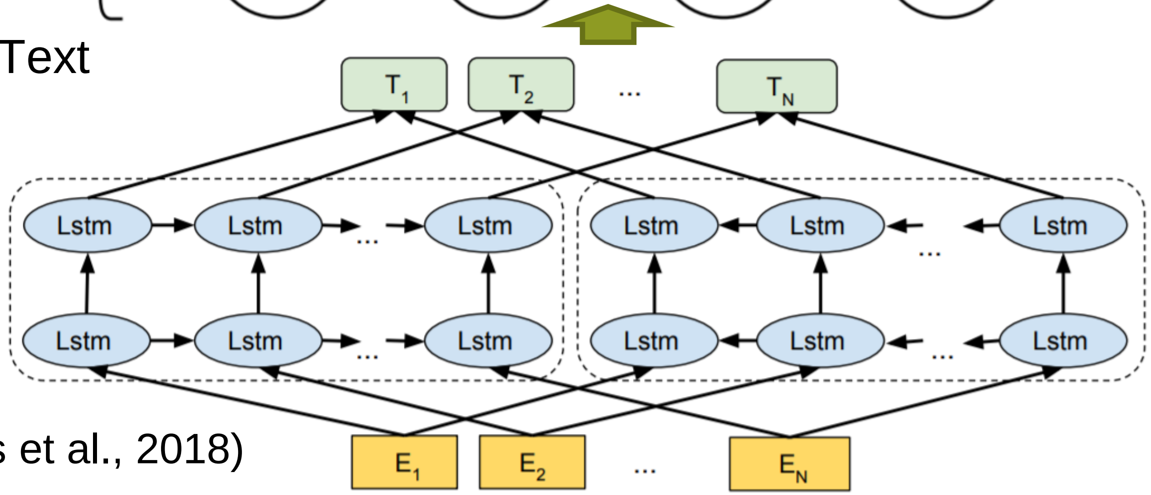


BERT (Devlin et al., 2019)

BiLSTM-CRF (Ma and Hovy, 2016)



fastText



ELMo (Peters et al., 2018)

Query Strategies

- MNLP:

Unannotated objects are sorted in ascending order by the average log probability of sequence tags

$$\mathrm{MNLP} = \max_{\{y_j\}} \frac{1}{n} \sum_{i}^{n} \log P(y_i | \{y_j\} \setminus y_i, \{x_j\})$$

- Modification MNLP-mod:

$$\text{MNLP-mod} = \text{MNLP} \cdot \alpha, \ \text{where}$$

$$\alpha = \begin{cases} \frac{1}{\gamma} & \text{if y contains a tag `B-<type>'} \\ 1 & \text{otherwise} \end{cases}$$

Skoltech
Skolkovo Institute of Science and Technology

Corpora for Experiments

- I2B2 Heart risk factors (Stubbs et al., 2014)
  - We generated three datasets with entity-level annotations using the original data with document-level annotations

	Hypertension	**CAD**	**Diabetes**
Train, # sent.	9,871	25,924	14,183
Test, # sent.	6,813	16,560	8,088
% with entities	13.0	3.5	7.3

- JNLPBA /Genia (Collier et al., 2004)
  - 18,546 sentences for training and 3,856 for testing
  - 5 types of entities: “DNA”, “protein”, “cell type”, and “cell line”

BERT Finetuning Details

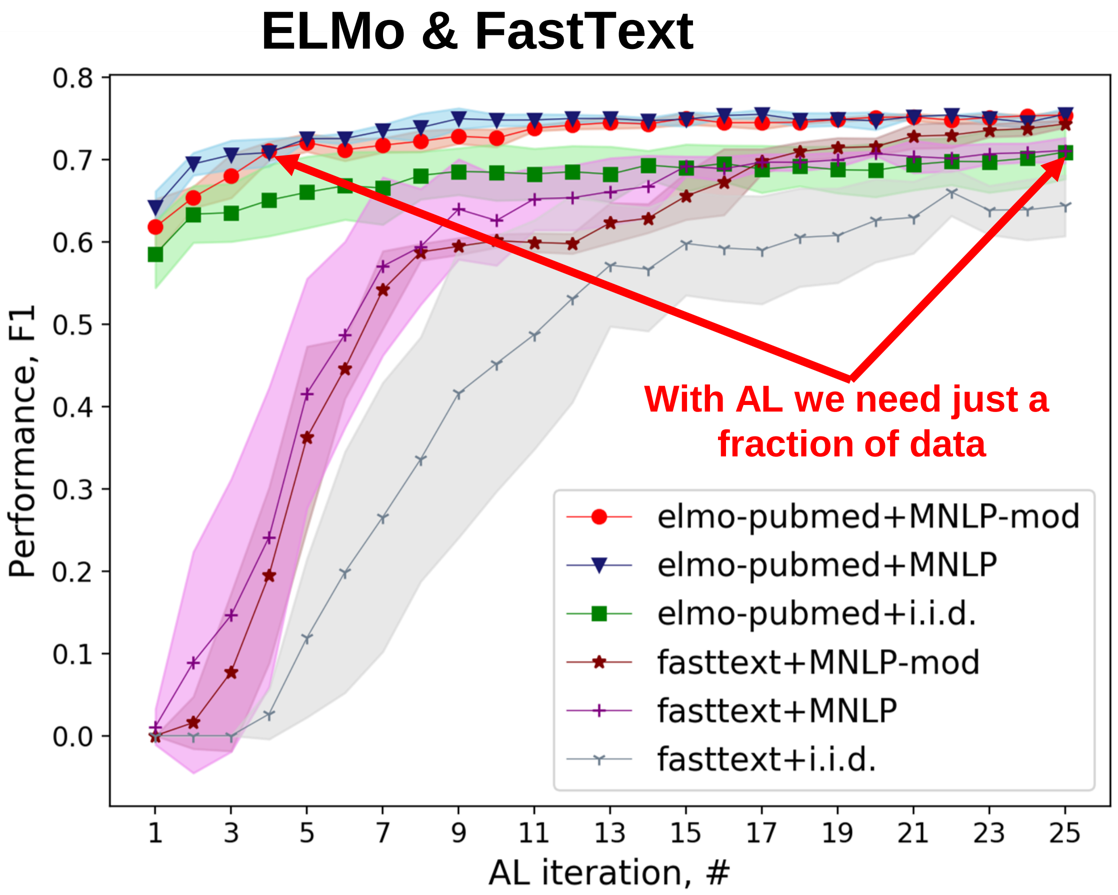
- You cannot finetune BERT like (Devlin, et al 2019) on very small data
- They use learning rate scheduler: warm-up over the first steps, and linear decay of the learning rate
- With very small data such scheduler is detrimental

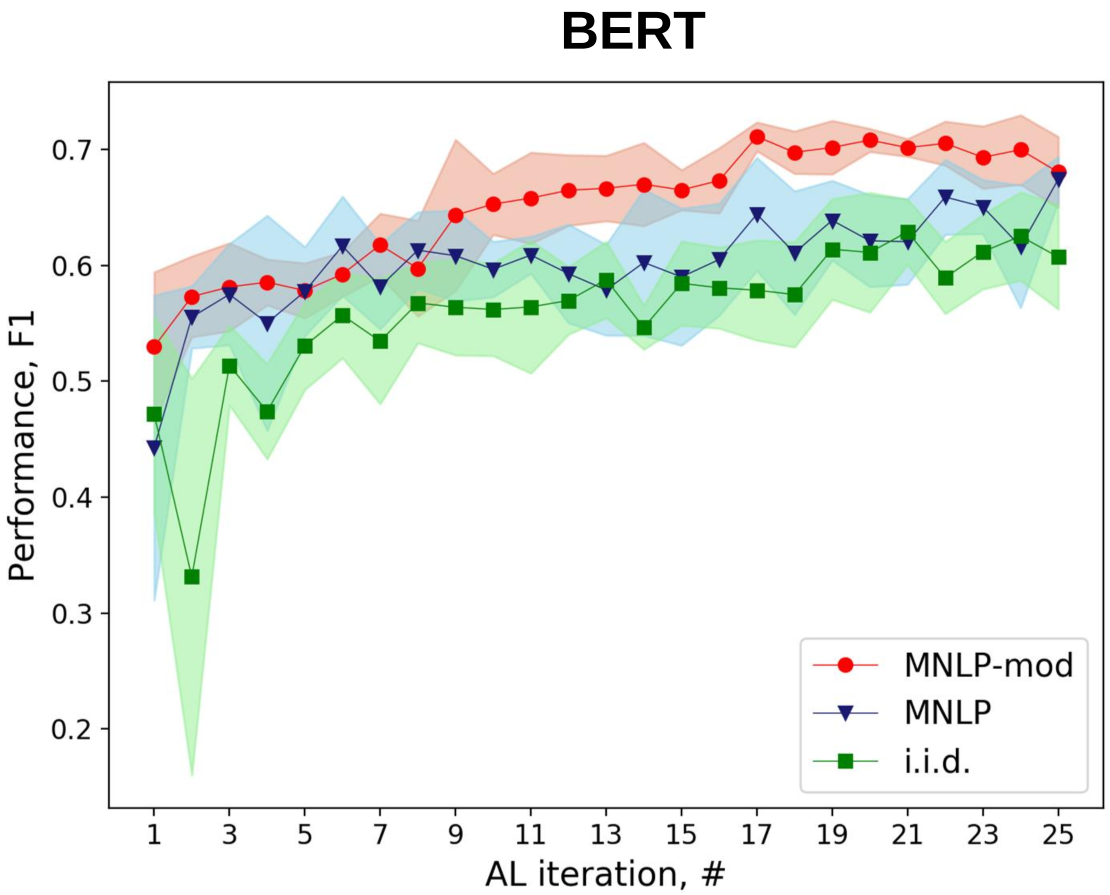
We used:

- Early stopping with number of tolerance epochs of 4, max number of epochs: 20 (however, in most cases BERT stops training earlier)
- Adam, learning rate: 5e-5 (*10 for the head), 0.01 L2 weight decay, batch size 45, gradient clipping: 1.0
- No learning rate annealing

Skoltech
Skolkovo Institute of Science and Technology

Results on i2b2 Heart Risk Factors (Diabetes)



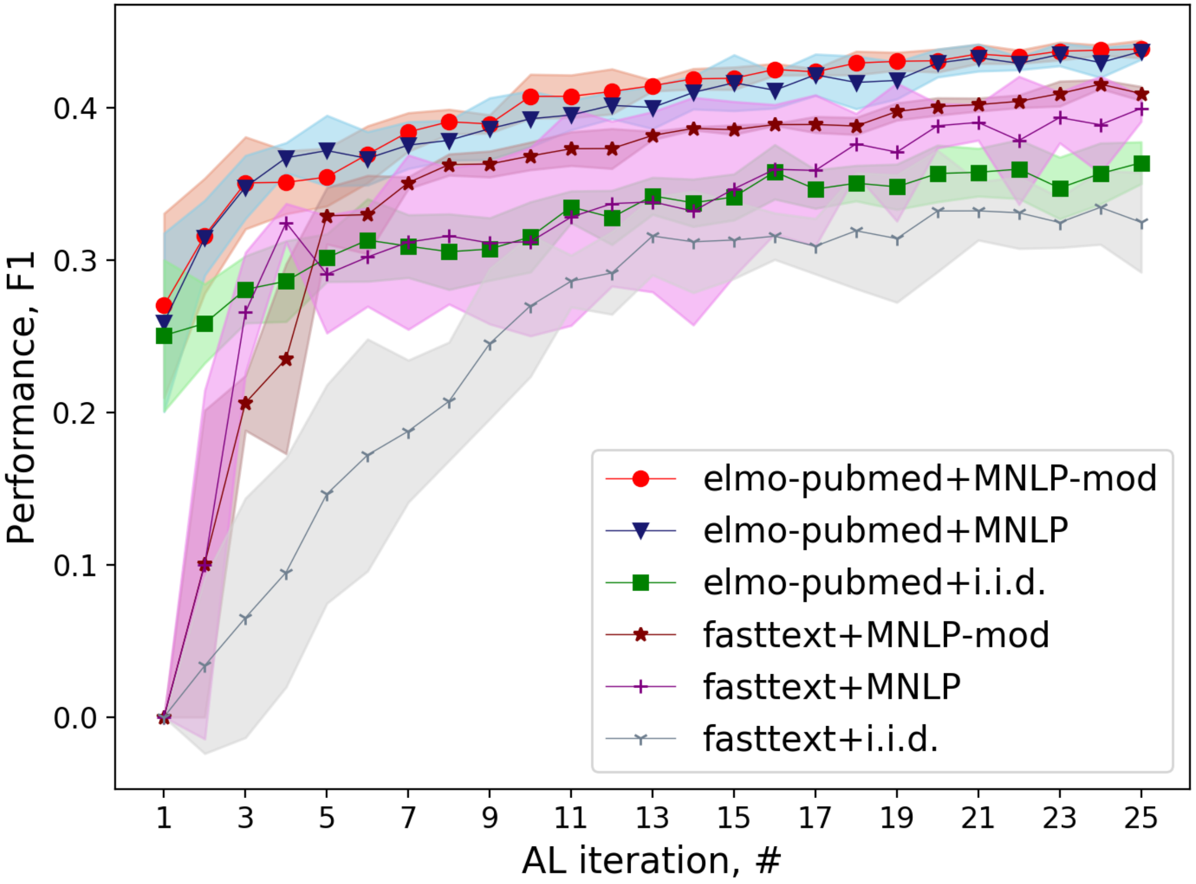


- Active learning is better than i.i.d. sampling on every dataset and with every model
- Sequence taggers based on deep pre-trained models can be trained on very small data compared to the model based on shallow DSM (fastText)

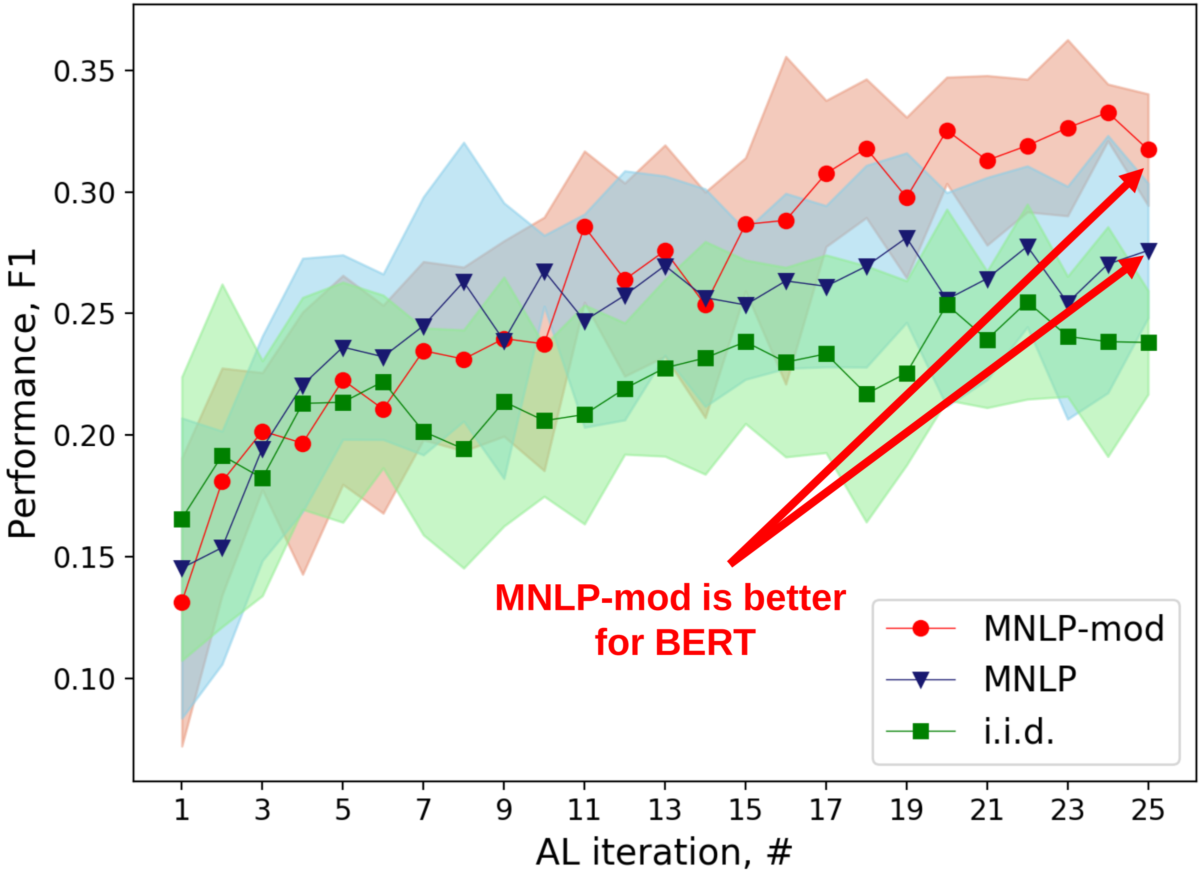
Skoltech
Skolkovo Institute of Science and Technology

Results on i2b2 Heart Risk Factors (CAD)

ELMo & FastText



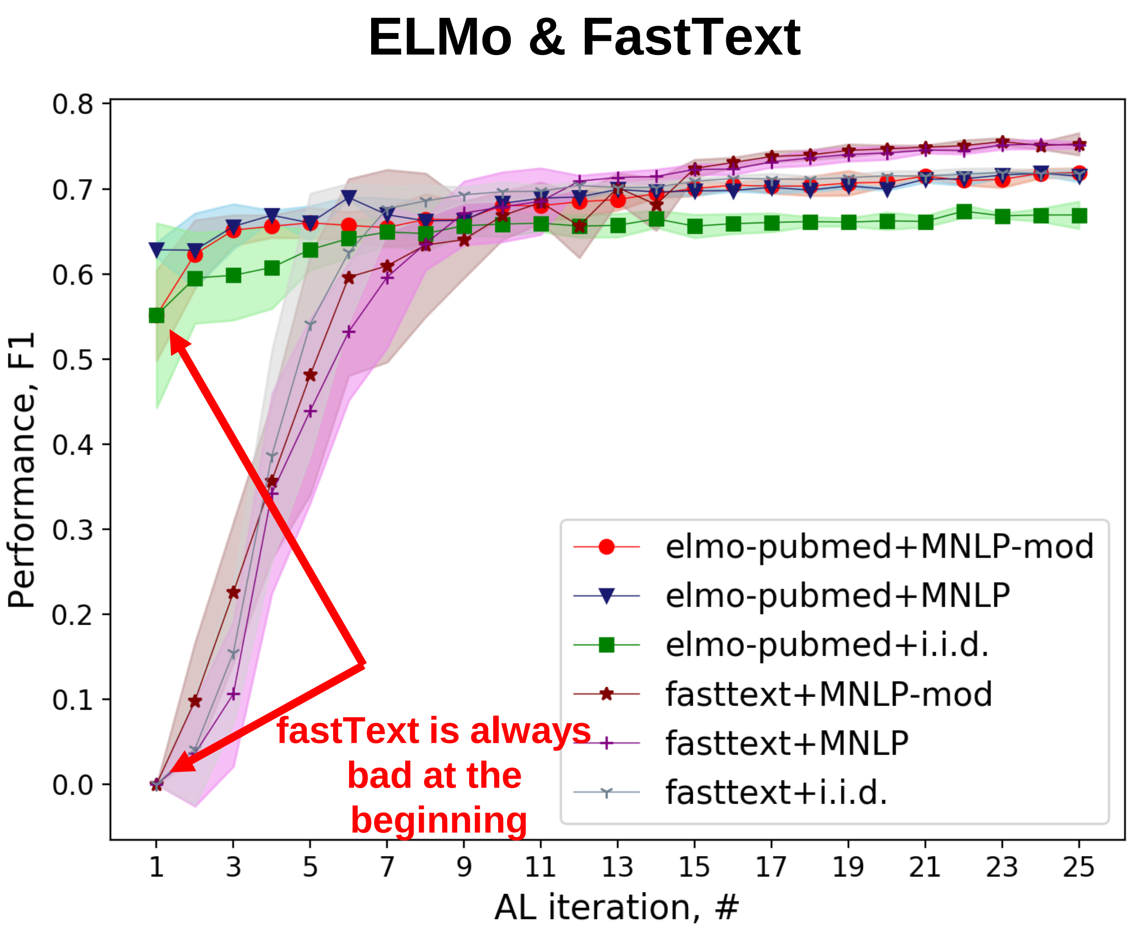
BERT

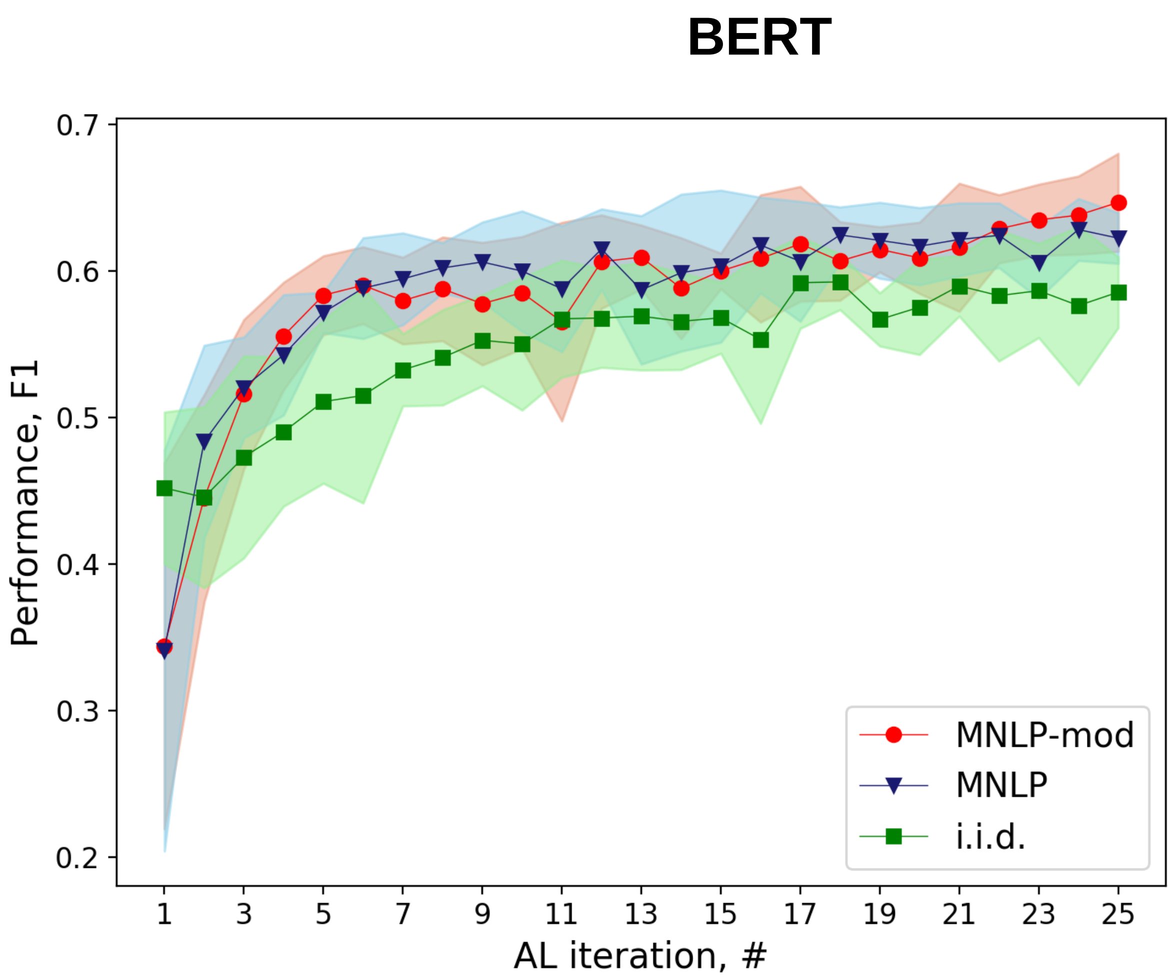


- MNLP-mod potentially helps to deal with very skewed datasets

Skoltech
Skolkovo Institute of Science and Technology

Results on i2b2 Heart Risk Factors (Hypertension)



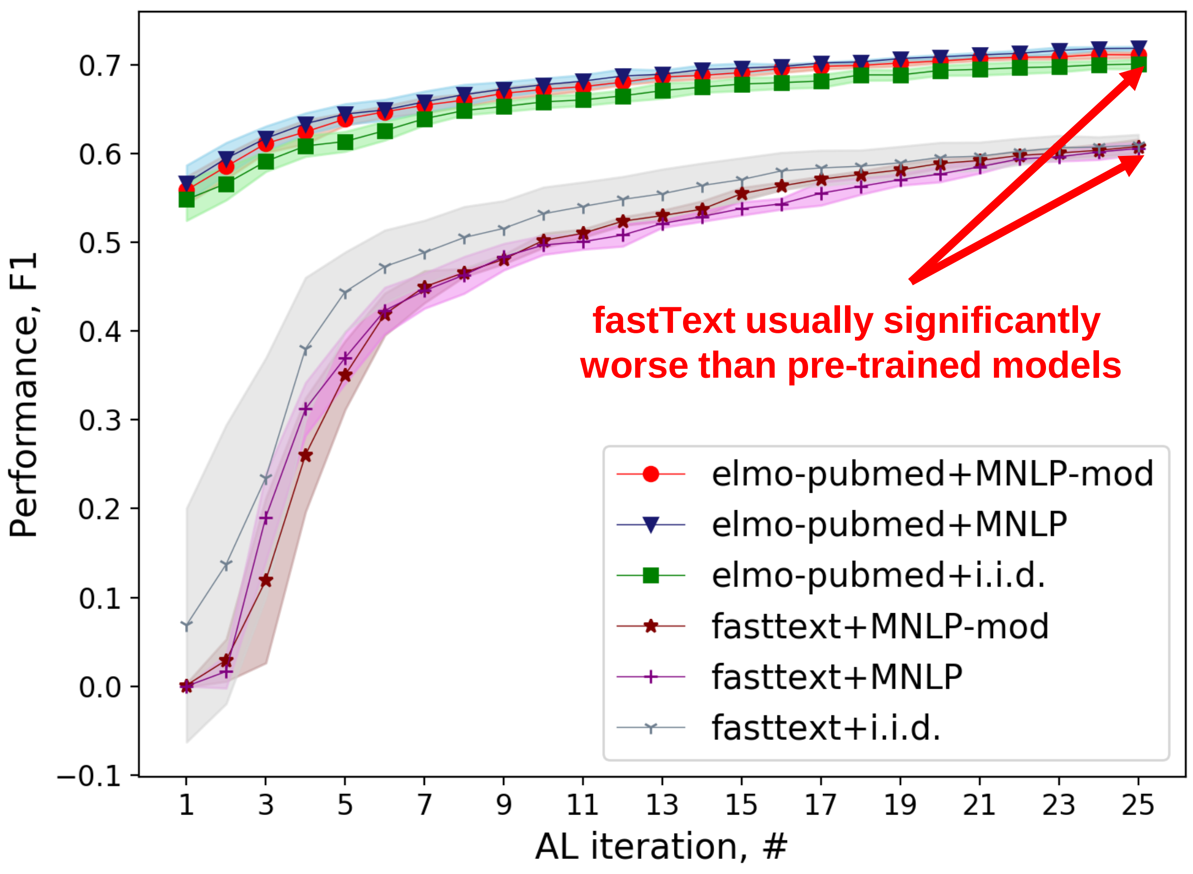


- In this experiment, fastText outperforms deep pre-trained models, although it still worse in the beginning

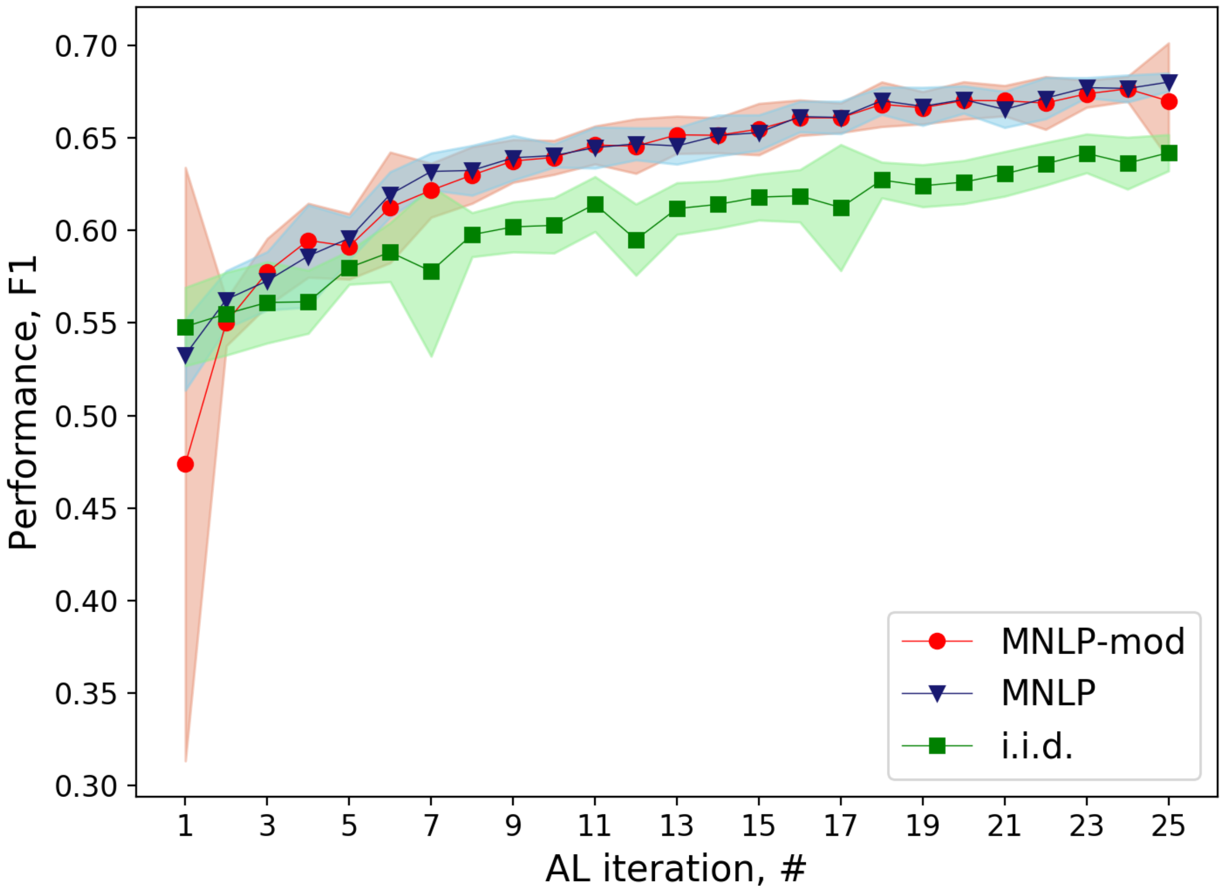
Skoltech
Skolkovo Institute of Science and Technology

Results on JNLPBA

ELMo & FastText



BERT



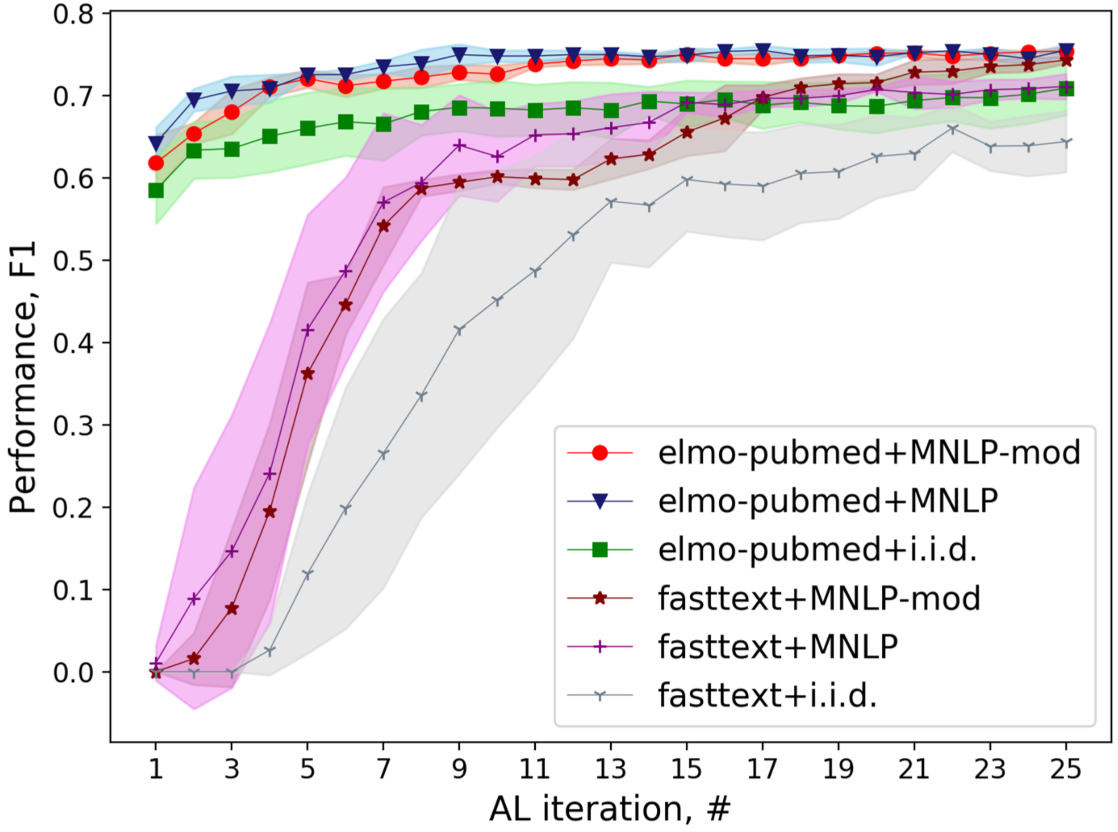
- Deep pre-trained models overall perform better than fastText (except hypertension dataset)

Skoltech
Skolkovo Institute of Science and Technology

Summary

- Active learning is better than i.i.d. sampling on every dataset and with every model
- Sequence taggers based on deep pre-trained models can be trained on very small data compared to the model based on shallow DSM
- Deep pre-trained models overall perform better than fastText (except hypertension dataset)
- ELMo has the best performance overall, but BERT is several times faster, so it is still practical to favor BERT in AL

ELMo & FastText



i2b2: Diabetes

Skoltech
Skolkovo Institute of Science and Technology

AL for Biomedical Research in Cardiology

In conjunction with
National Cardiological Center

Skoltech
Skolkovo Institute of Science and Technology

We use AL for Biomedical Research in Cardiology

Ишемическая болезнь сердца Артериальная гипертония

Хроническая сердечная недостаточность Сахарный диабет

Фибрилляция предсердий

Диагноз заключительный ИБС: Инфаркт миокарда без подъема сегмента ST от 05.01.18г. Ранняя постинфарктная стенокардия. Транслюминальная балонная ангиопластика коронарных артерий со стентированием ствола левой коронарной артерии с переходом на проксимальный и средний сегмент передней нисходящей артерии стентами Promus Element 4,0x32мм и Promus Element 3,5x38мм., проксимальной трети от устья огибающей артерии Promus Element 3,5x12 мм. от 18.01.18г. Атеросклероз коронарных артерий (окклюзия ПКА, субтотальный стеноз ствола ЛКА, 90% стеноз устья ОА). Постинфарктный кардиосклероз (инфаркт миокарда от 2004г).Нарушение ритма сердца: впервые возникший пароксизм фибрилляции предсердий, тахиформа от 15.01.18г. Впервые возникший пароксизм трепетания предсердий от 18.01.18г. Хроническая сердечная недостаточность 2ФК по NYHA. Артериальная гипертензия 3 ст, риск 4. Сахарный диабет 2 типа. Диабетическая микромакроангиопатия. Диабетическая дистальная полинейропатия, сенсорно-моторная форма.Синдром диабетической стопы, нейроишемическая форма.Облитерирующий атеросклероз нижних конечностей. Балонная ангиопластика и стентирование левой ПБА от 19.05.11г.

Ischemic stroke risk assessment:

CHA2DS2-VASc:

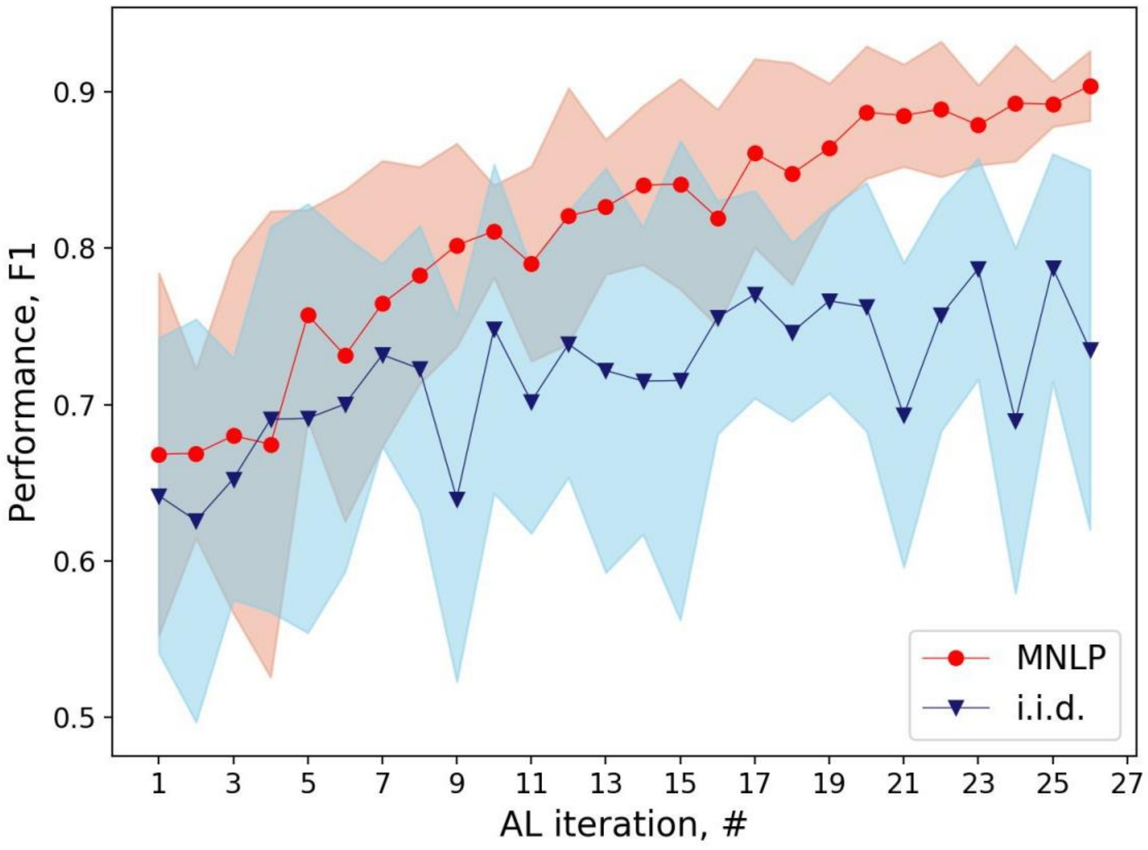
4 пунктов

Skoltech

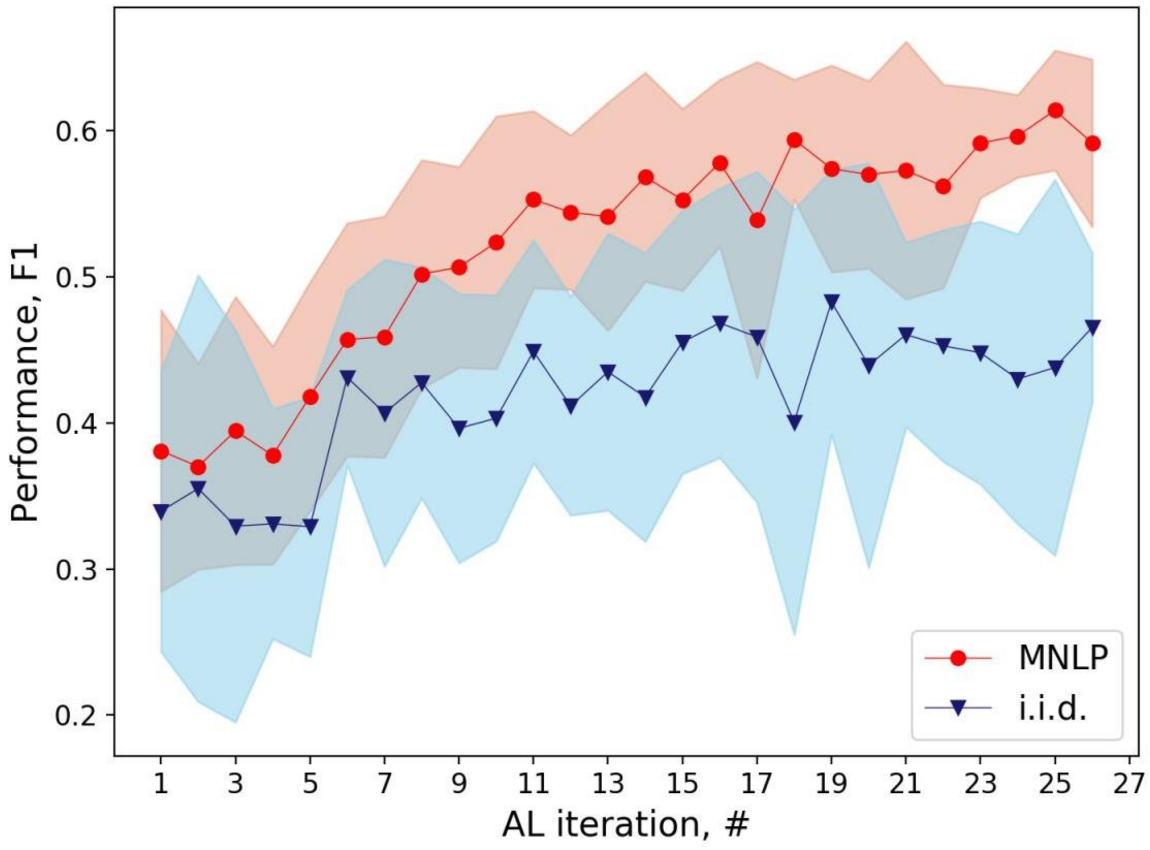
Skolkovo Institute of Science and Technology

Results on Russian-language Data from National Cardiological Center (1)

Hypertension



Peripheral Arterial Disease

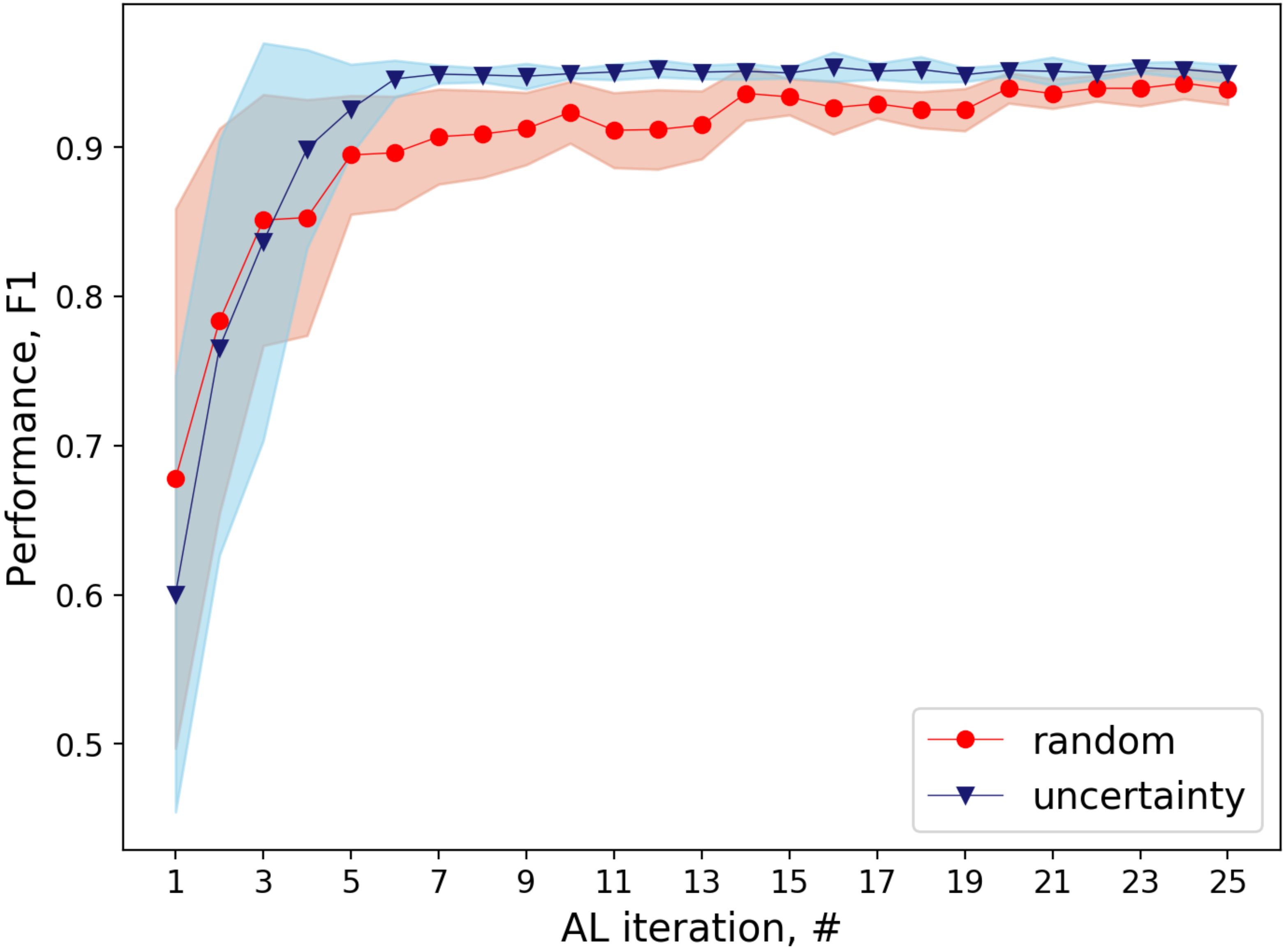


BERT for token classification (based on RuBERT)

Skoltech

Skolkovo Institute of Science and Technology

Results on Russian-language Data from National Cardiological Center (2)



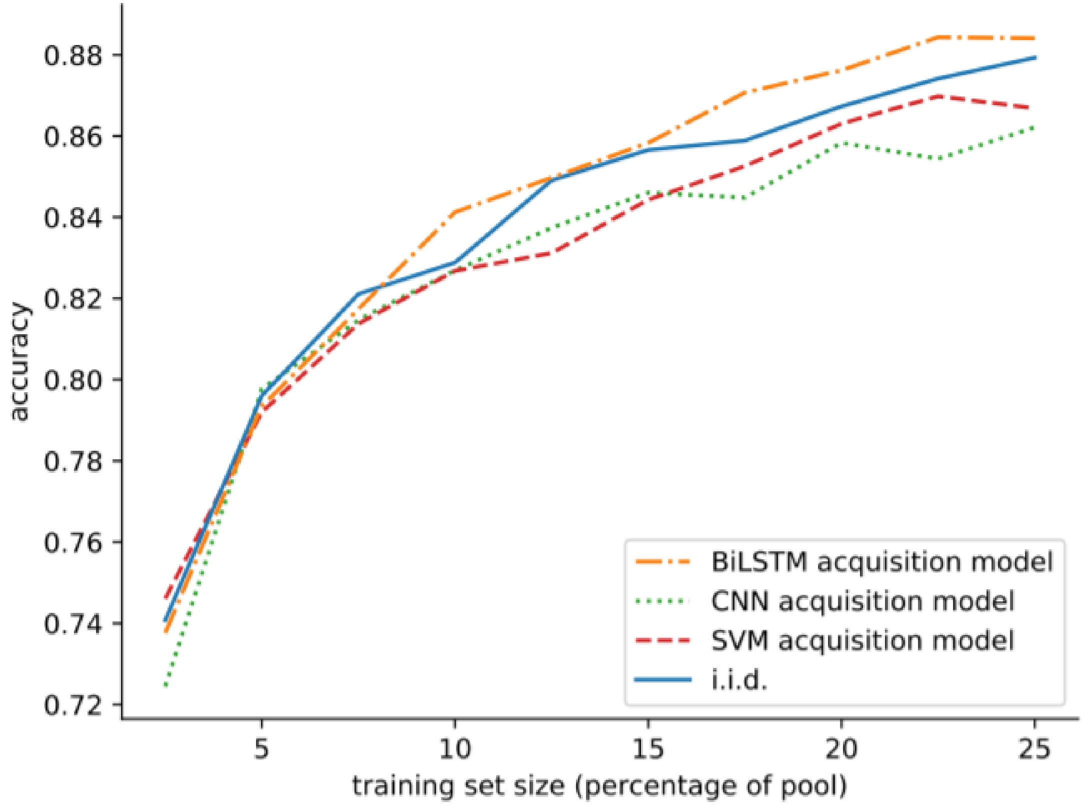
- Hypertension
- ELMo + BiLSTM-CRF
- ELMo for Russian from RusVectores

Skoltech
Skolkovo Institute of Science and Technology

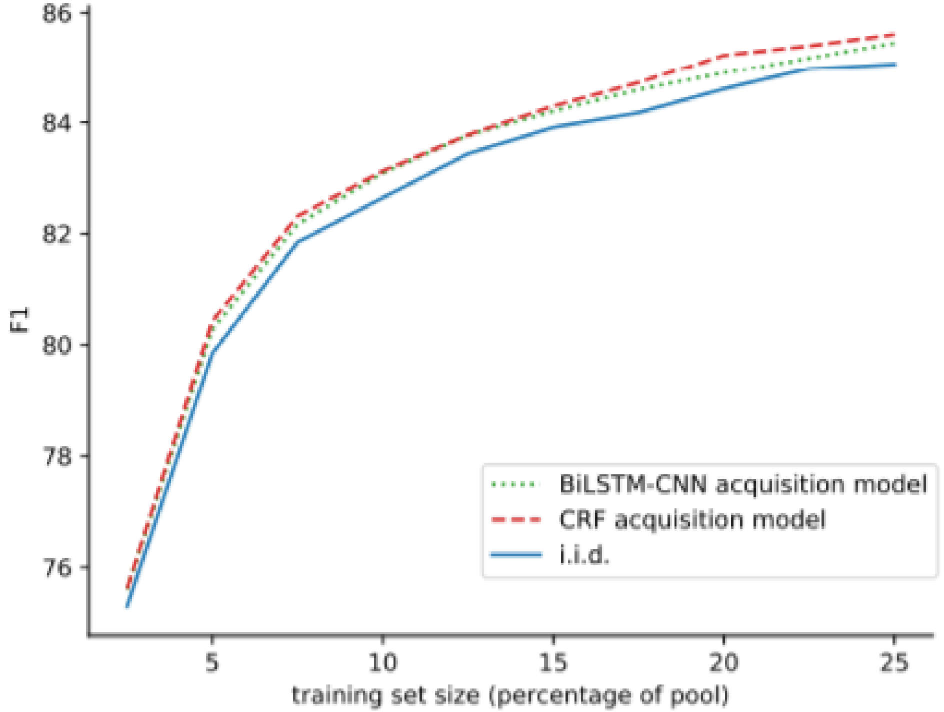
Disclaimer: AL sometimes does not work!

EMNLP 2019: "Practical obstacles to deploying active learning" (Lowell et al., 2019)

→ If you use one model to create a dataset with AL and train another model on the result dataset you can get a performance drop!



BiLSTM performance on text classification
Subjectivity corpus (Pang and Lee, 2004)



BiLSTM-CNN on
OntoNotes 5.0

Skoltech
Skolkovo Institute of Science and Technology

Key Takeaways

- **Do not write hand-crafted rules! Instead, annotate quickly!**
- **Deep pre-trained models and active learning is a powerful combination**

- Active learning is especially good when you cannot do crowdsourcing (e.g., in clinical medicine or biomedicine)
- BERT training procedure on very small data is different from the method presented in the original paper (Devlin et al., 2019)
- BERT performed worse in the AL setting (in our experiments) than ELMo-BiLSTM-CRF. However, it is computationally faster
- AL is biased sampling a priory! You cannot test on such data
- AL sometimes does not work! Especially when you use different models for acquisition and evaluation

References (1)

- A. Culotta and A. McCallum. 2005. Reducing labeling effort for stuctured prediction tasks. In Proceedings of the National Conference on Artificial Intelligence (AAAI) , pages 746–751. AAAI Press.
- Erdmann, Alexander, et al. "Challenges and solutions for Latin named entity recognition." Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). 2016.
- Erdmann, Alexander, et al. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.
- Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning //international conference on machine learning. – 2016. – P. 1050-1059.
- Gal Y., Islam R., Ghahramani Z. Deep Bayesian active learning with image data //Proceedings of the 34th International Conference on Machine Learning-Volume 70. – JMLR. org, 2017. – P. 1183-1192.

Skoltech
Skolkovo Institute of Science and Technology

References (2)

- I. Dagan and S. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In Proceedings of the International Conference on Machine Learning (ICML) , pages 150–157. Morgan Kaufmann
- Lowell, David, Zachary C. Lipton, and Byron C. Wallace. "How transferable are the datasets collected by active learners?." arXiv preprint arXiv:1807.04801 (2018).
- Ma X., Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2016.
- McCallum A., Nigam. K. 1998. Employing EM in pool-based active learning for text classification. In Proceedings of the International Conference on Machine Learning (ICML) , pages 359–367
- S. Kim, Y. Song, K. Kim, J.W. Cha, and G.G. Lee. 2006. MMR-based active machine learning for bio named entity recognition. In Proceedings of Human Language Technology and the North American Association for Computational Linguistics (HLT-NAACL) , pages 69–72. ACL Press.

Skoltech
Skolkovo Institute of Science and Technology

References (3)

- Settles, Burr, and Mark Craven. "An analysis of active learning strategies for sequence labeling tasks." Proceedings of the conference on empirical methods in natural language processing.
- Shen, Yanyao, et al. "Deep Active Learning for Named Entity Recognition." ICLR. 2018.
- Siddhant, Aditya, and Zachary C. Lipton. "Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.
- Suvorov, Roman, Artem Shelmanov, and Ivan Smirnov. "Active Learning with Adaptive Density Weighted Sampling for Information Extraction from Scientific Papers." Conference on Artificial Intelligence and Natural Language. Springer, 2017.
- T. Scheffer, C. Decomain, and S. Wrobel. 2001. Active hidden Markov models for information extraction. In Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA) , pages 309–318. Springer-Verlag.