

Intro to NLP

Katya Artemova

Computational Pragmatics Lab, HSE

September 2, 2019

Today

Intro

About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

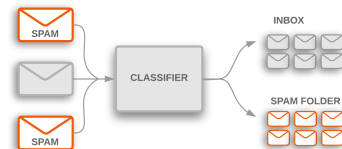
Natural language processing ...

- ▶ along with computer vision a crucial part of modern artificial intelligence
- ▶ deals with all human (and machine) interactions in language
- ▶ requires understanding of linear algebra, statistics, mathematics in general, linguistics and coding skills

Example tasks

Text classification

- ▶ Sentiment analysis
- ▶ Intent detection
- ▶ Spam filtering
- ▶ Topic classification



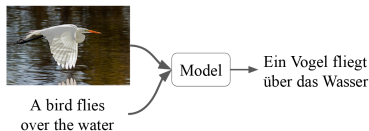
Sequence labelling

- ▶ Named entity recognition
- ▶ Coreference resolution

contentShip to site index? Policies Subscribed Log In Today's [Paper Advertisements Supported](#) [066](#) by? B.I. Agent [Peter Strick](#) [PERSON](#).
[Who Criticized Trump](#) [PERSON](#) = Tests, in [Fandango](#) [Peter Strick](#), a top [F.B.I. SFE](#) counterintelligence agent who was taken off the special counsel
investigation after his disparaging tweets about President [Trump](#) [PERSON](#) were uncovered, was fired. [Credit J. Kirkpatrick](#) [PERSON](#) for [The New York](#)
[Times](#) [Adam Goldman](#) [CNN](#) and [Michael S. Schmidt](#) [NY](#) [PERSON](#) 13 [CANDIDATE](#) 2018 WASHINGTON [CANDIDATE](#) — [Peter Strick](#)
[PERSON](#) The [F.B.I. SFE](#) senior counterintelligence agent who disparaged President [Trump](#) [PERSON](#) in inflammatory text messages and helped
oversee the [Hillary Clinton](#) [PERSON](#) email and [Russia](#) [SFE](#) investigations, has been fired for violating bureau policies, [NY](#) [Strick](#) [PERSON](#)'s lawyer
said [Monday](#) [DATE](#) [NY](#) [Strick](#) and his allies seized on the tweets — exchanged during the [2015](#) [DATE](#) campaign with a former [F.B.I. SFE](#) lawyer,
[Lisa Page](#) — [NY](#) [PERSON](#) enclosing the [Russia](#) [SFE](#) investigation as an illegitimate "witch hunt." [Strick](#) [PERSON](#) who rose over [20](#) years
[DATE](#) at the [F.B.I. SFE](#) to become one of its most experienced counterintelligence agents, was a key figure in the early months [DATE](#) of the
inquiry. Along with writing the tweets, [Strick](#) [PERSON](#) was accused of sending a highly sensitive search warrant to his personal email account. The
[F.B.I. SFE](#) had been under immense political pressure by [NY](#) [Trump](#) [PERSON](#) to dismiss [Strick](#) [PERSON](#) who was removed last summer
[DATE](#) from the staff of the special counsel, [Robert S. Mueller Jr.](#) [PERSON](#) The president has repeatedly denounced [NY](#) [Strick](#) [PERSON](#) as posts on

Sequence transformation (seq2seq)

- ▶ Machine translation
- ▶ Question answering



Phenomena to handle

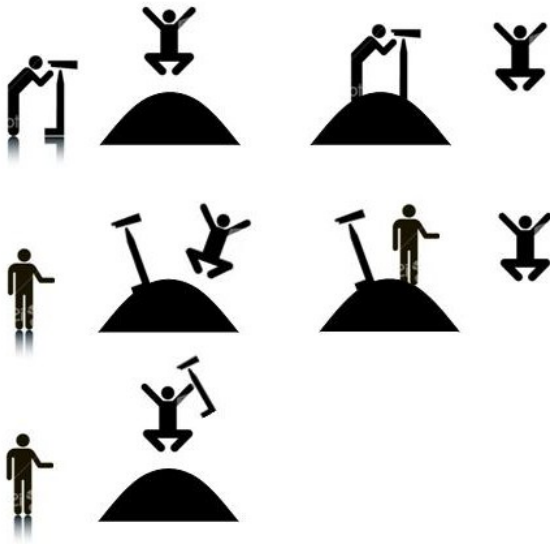
1. Tokenization and sentence boundary detection
2. Morphology
3. Syntax
4. Semantics
5. Multilinguality

Ambiguity

1. Polysemy and word-sense disambiguation: , bank
2. Homonymy: the ship or to ship,
3. Syntactic ambiguity: John saw the man on the mountain with a telescope.

Syntactic ambiguity

John saw the man on the mountain with a telescope



Today

Intro

About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

About this course

- ▶ **Team:** Katya Artemova (lectures), Veronika Sarkisyan (seminars), Vadim Fomin (TA, Sberbank)
- ▶ **Repo:** github.com/PragmaticsLab/NLP-course-FinTech
- ▶ **Chat:** t.me/nlp_fintech
- ▶ **Final mark:**
$$M_{1,2} = \text{round}(0.2\text{quiz} + 0.5\text{HW} + 0.3\text{project})$$
$$\text{final} = \text{round}(0.4\text{exam} + 0.3(M_1 + M_2))$$
- ▶ **Project:** SemEval or similar shared tasks:

Our plan

1. Word embeddings
2. Text classification
3. Sequence modelling
4. Walk down Sesame Street
5. Syntax
6. Machine translation
7. Natural language generation

Today

Intro

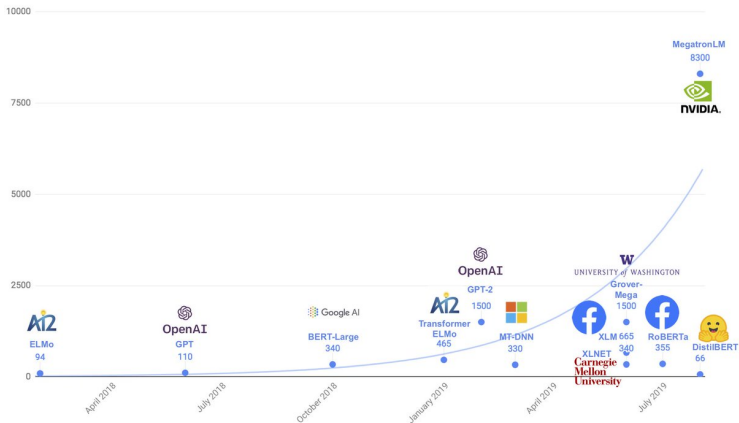
About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

NLP's ImageNet moment has arrived



... but is rather questionable

Recent trends in NLP

1. **The ethics of AI**

- ▶ Fairness
- ▶ Societal applications

2. **Transfer learning**

- ▶ Cross-lingual methods
- ▶ Cross-domain methods

3. **Question answering**

4. **Multimodal NLP**

5. **Clinical NLP**

Today

Intro

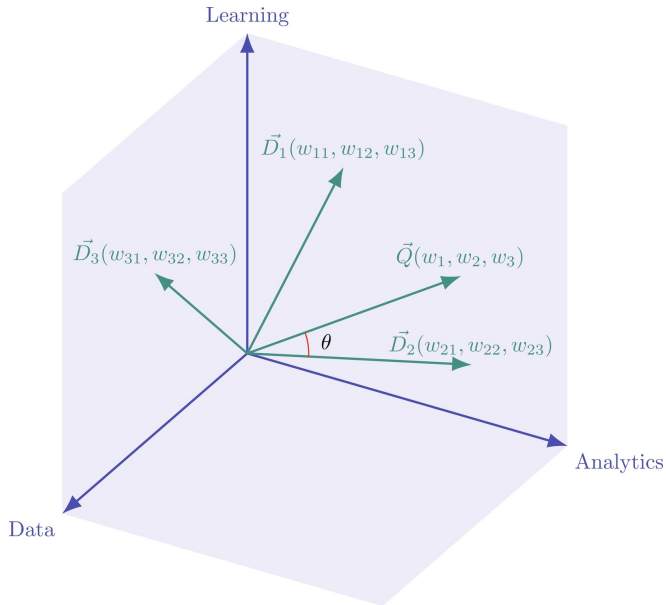
About this course

Recent trends in NLP

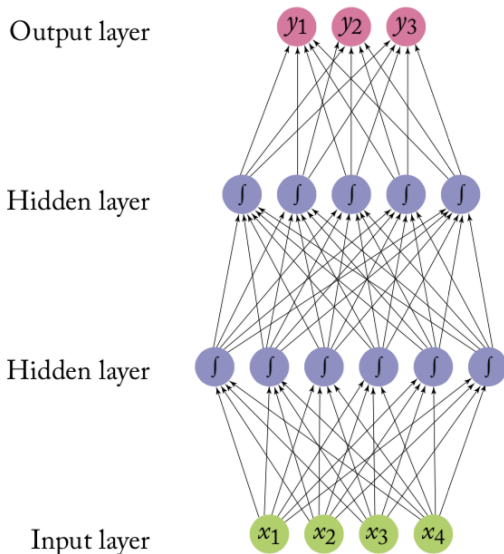
Example task: text classification

Practice: tools for processing Russian

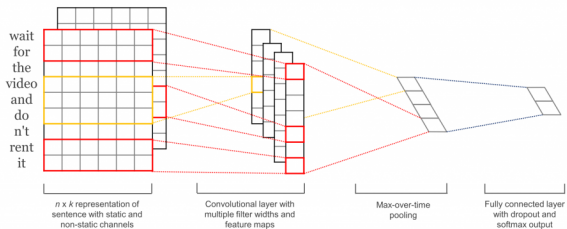
Vector space model [1]



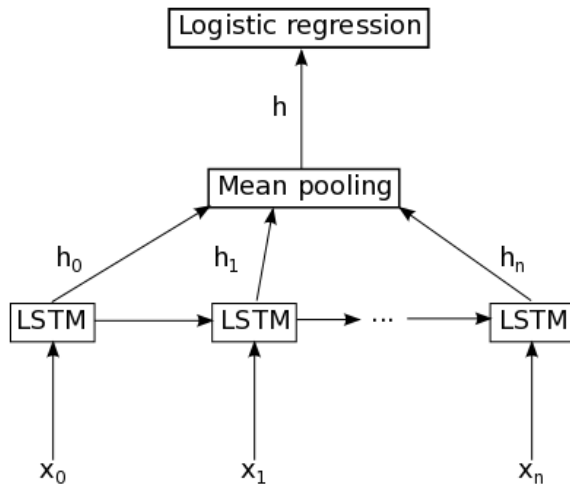
Feed forward network



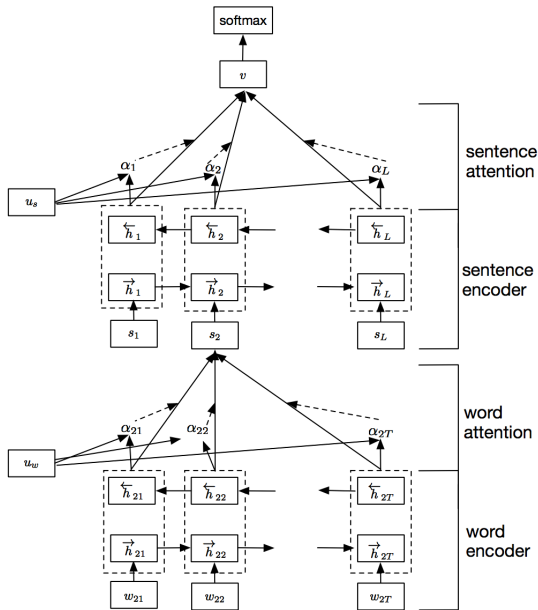
Convolutional network [2]



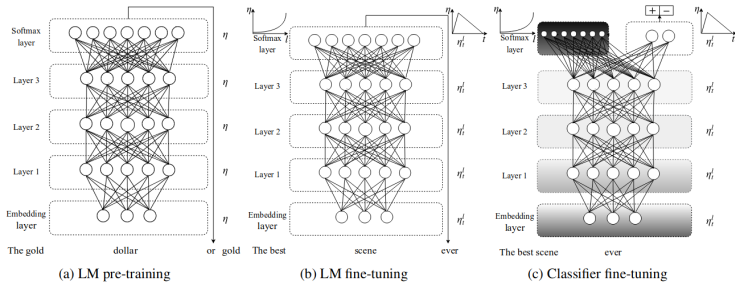
LSTM



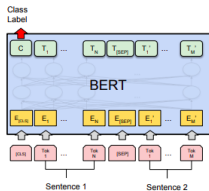
Hierarchical attention network [3]



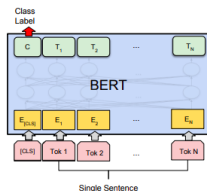
ULMFiT [4]



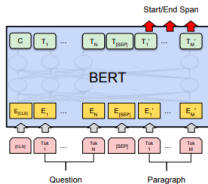
BERT [5]



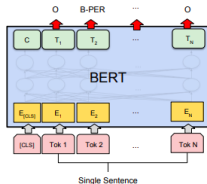
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Today

Intro

About this course

Recent trends in NLP






Example task: text classification

Practice: tools for processing Russian

Reading

1. Text classification algorithms: a survey [arXiv]
2. Speech and Language Processing. Daniel Jurafsky, James H. Martin, Ch. 2 [url]
3. Natural Language Processing. Jacob Eisenstein, Ch. 2-4, [[GitHub]

Reference

-  G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
-  Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.
-  Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.
-  J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” *arXiv preprint arXiv:1801.06146*, 2018.
-  J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.