

Intro to NLP

Dmitry Ilvovsky and Ekaterina Chernyak

September 1, 2020

Today

Intro

About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

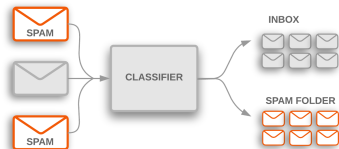
Natural language processing ...

- ▶ 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



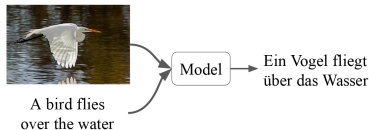
Sequence labelling

- ▶ Named entity recognition
- ▶ Coreference resolution

contentShip to site index? Politics Subscribed Log In Today's [Paper](#) Advertisements Supported [066](#) by? B.I. Agent [Peter Strick](#) [PERSON](#).
[Who Criticized Trump](#) [PERSON](#) in Texas, in [Fandango](#) Peter Strick, a top [F.B.I. SFE](#) counterintelligence agent who was taken off the special counter
investigation after his disparaging texts 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 [CARDINAL](#) 2018 WASHINGTON [CARDINAL](#) — [Peter Strick](#)
[PERSON](#) the [F.B.I. SFE](#) senior counterintelligence agent who disparaged President [Trump](#) [PERSON](#) in inflammatory text messages and helped
coerce the [Henry Clinton](#) [PERSON](#) email and [Raisa SFE](#) investigations, has been fired for violating bureau policies, Mr. [Strick](#) [PERSON](#)'s lawyer
said [Monday](#) [DATE](#) Mr. Trump and his allies seized on the texts — exchanged during the [2015](#) [DATE](#) campaign with a former [F.B.I. SFE](#) lawyer,
[Lisa Page](#) — [F.B.I. SFE](#) encoding the [Russian SFE](#) investigation as an illegitimate "witch hunt." Mr. [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 texts, Mr. [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 Mr. [Trump](#) [PERSON](#) to dismiss Mr. [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 Mr. [Strick](#) [PERSON](#) in posts on

Sequence transformation (seq2seq)

- ▶ Machine translation
- ▶ Question answering



Phenomena to handle

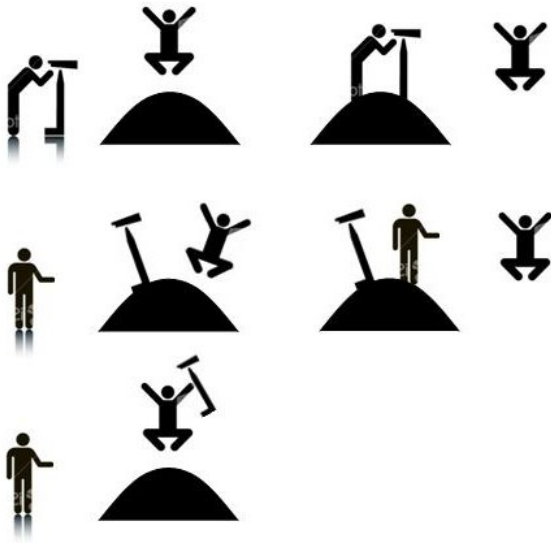
1. Tokenization and sentence boundary detection
2. Morphology
3. Syntax
4. Semantics
5. Discourse
6. Pragmatics
7. Multilinguality

Ambiguity

1. Polysemy and word-sense disambiguation: орган, bank
2. Homonymy: the ship or to ship, стекло
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

This course is based on the NLP course developed by Ekaterina Chernyak from parallel education program:

github.com/PragmaticsLab/NLP-course-AMI

- 1 Lecturer: Dmitry Ilvovsky
- 2 Seminars: Dmitry Ilvovsky, Denis Smirnov
- 3 TA: Denis Smirnov, denis.m.smirnov@gmail.com
- 4 Repo: github.com/denis-smirnov/hse-ami-nlp-course-fall-20
- 5 Chat: <https://t.me/joinchat/CDDAm03aoX53xIQi96JPoQ>
- 6 Final mark: $M_1, 2 = \text{round}(0,6HW + 0,4Project)$
 $final = \text{round}(0,4exam + 0,3(M_1 + M_2) + 0,5_{questions})$
- 7 Project: SemEval 2020 or similar shared task

Our plan

- 1 Word embeddings
- 2 Text classification
- 3 Sequence modelling
- 4 Seq2Seq modelling
- 5 Syntax
- 6 Machine translation
- 7 Generative models
- 8 Linguistic resources
- 9 Discourse and Argumentation

Today

Intro

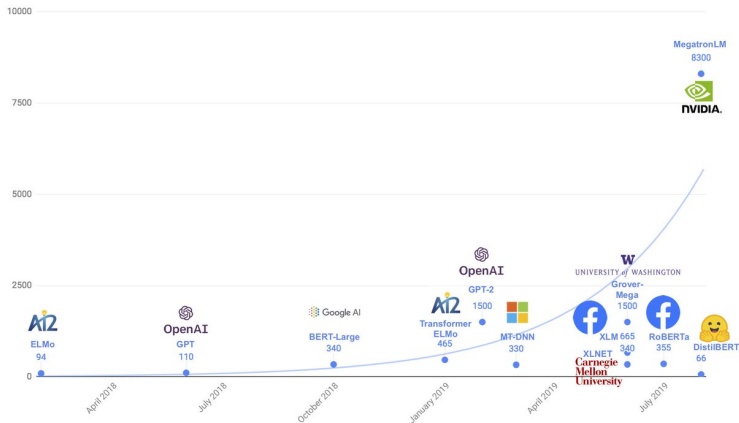
About this course

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

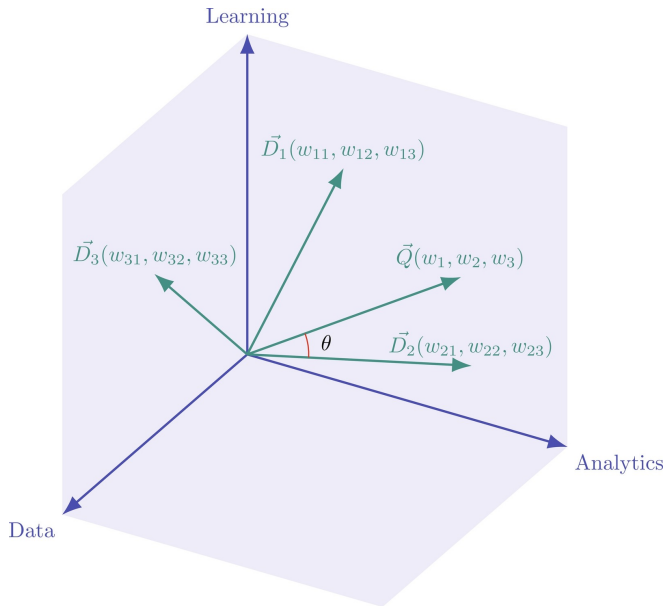
About this course

Recent trends in NLP

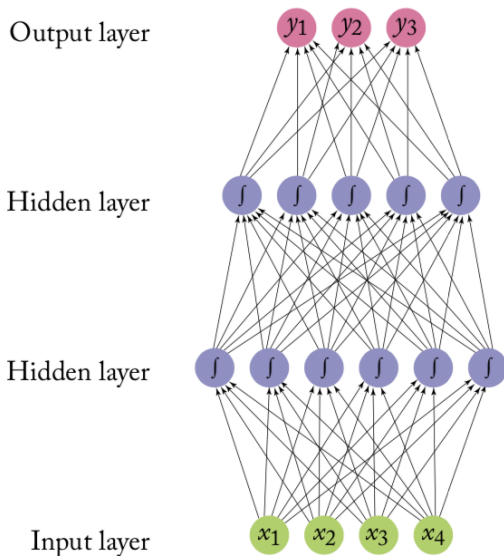
Example task: text classification

Practice: tools for processing Russian

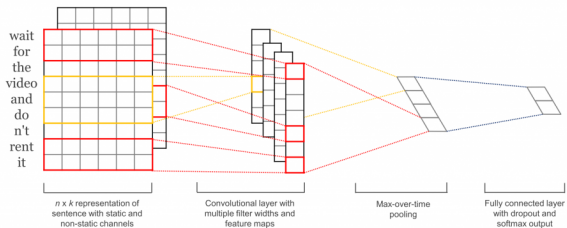
Vector space model [1]



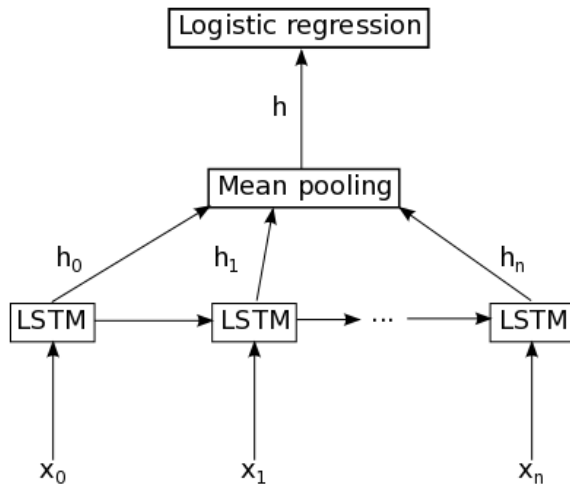
Feed forward network



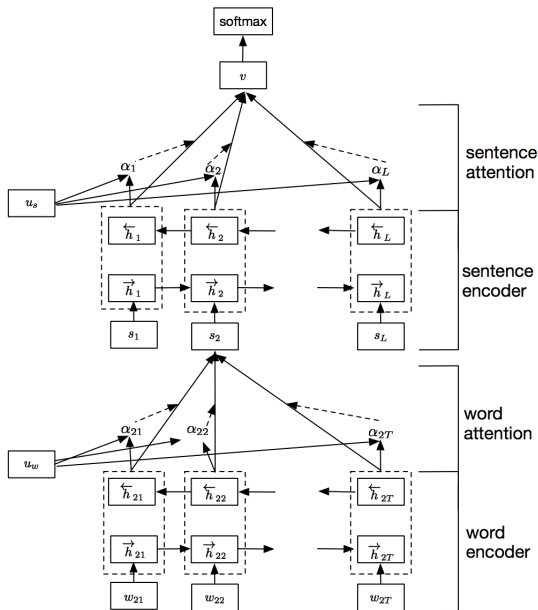
Convolutional network [2]



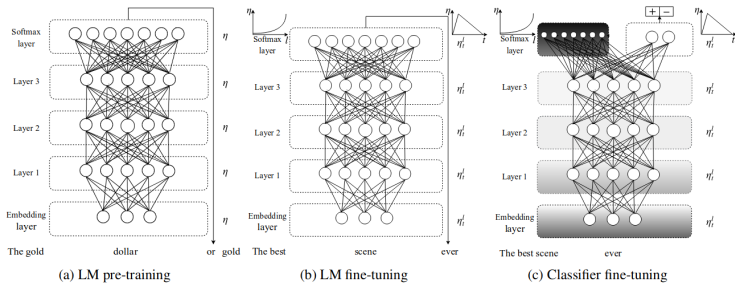
LSTM



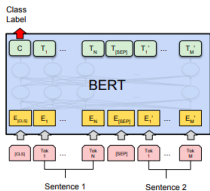
Hierarchical attention network [3]



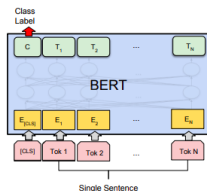
ULMFiT [4]



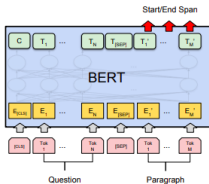
BERT [5]



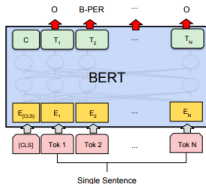
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(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 и C.-S. Yang, “A vector space model for automatic indexing”, *Communications of the ACM*, т. 18, № 11, с. 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 и E. Hovy, “Hierarchical attention networks for document classification”, в *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, с. 1480—1489.
-  J. Howard и S. Ruder, “Universal language model fine-tuning for text classification”, *arXiv preprint arXiv:1801.06146*, 2018.
-  J. Devlin, M.-W. Chang, K. Lee и K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*, 2018.