Intro to NLP

Katya Artemova

Computational Pragmatics Lab, HSE

4 сентября 2019 г.

Today

Intro

About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

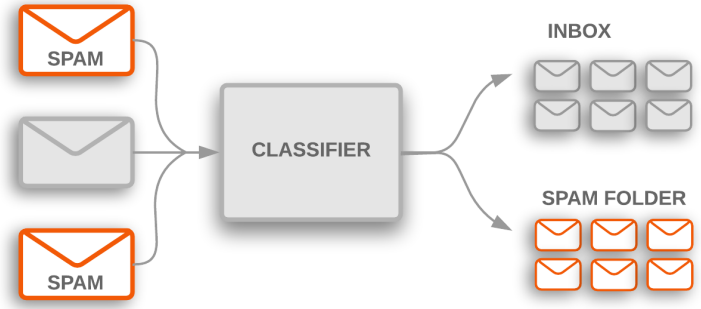
Natural language processing ...

- 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

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE . Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE 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. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Sequence transformation (seq2seq)

- Machine translation
- Question answering

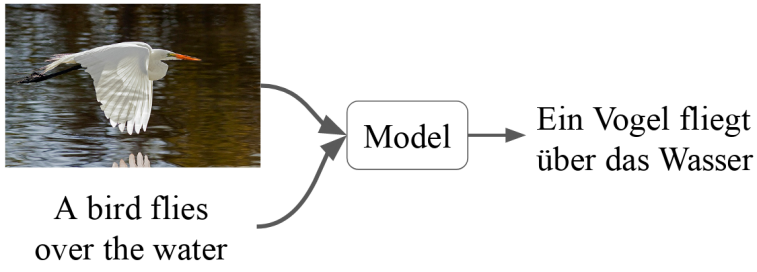


Image sources: developers.google.com, towardsdatascience.com, statmt.org

Phenomena to handle

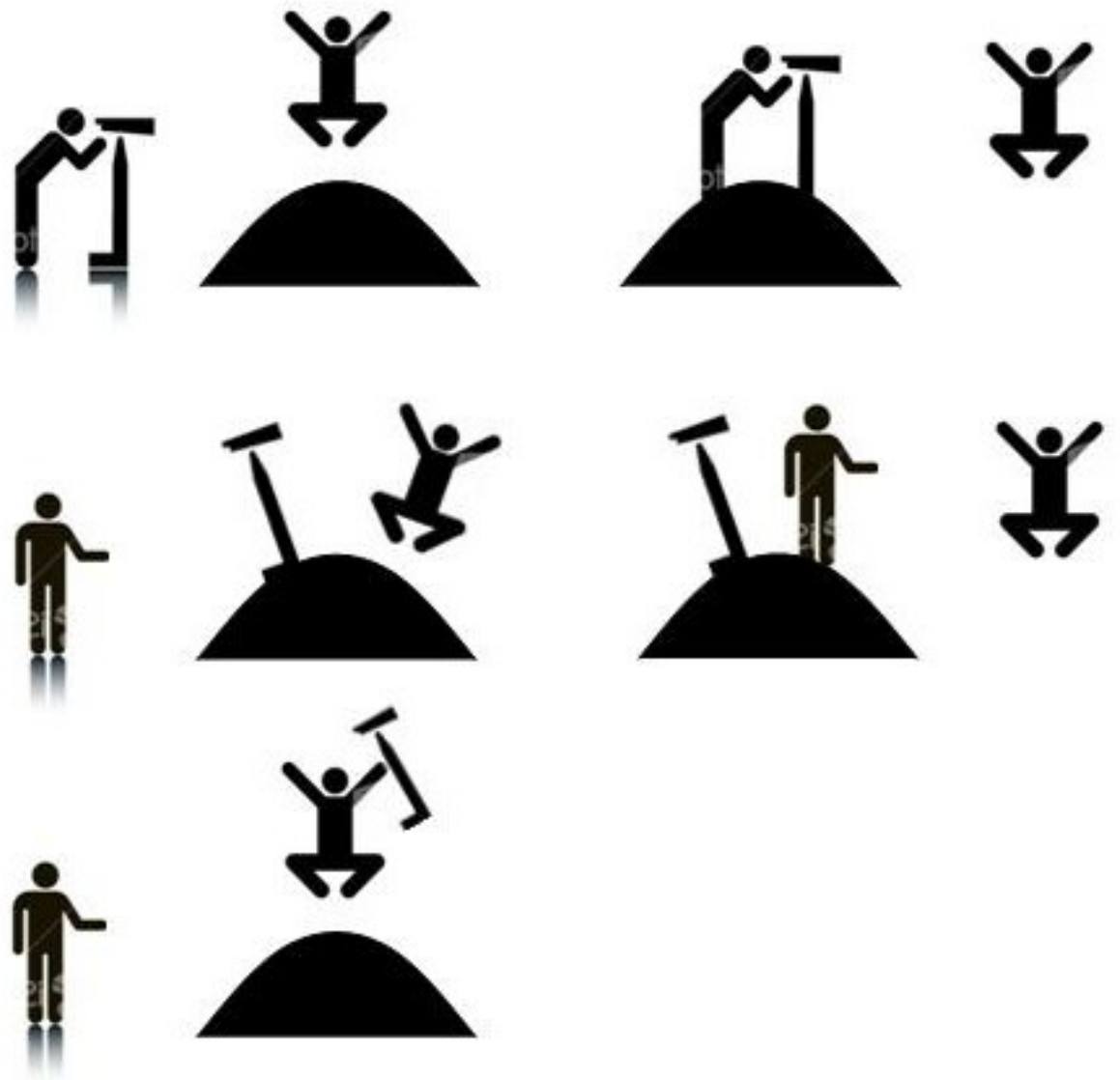
1. Tokenization and sentence boundary detection
2. Morphology
3. Syntax
4. Semantics
5. Discourse
6. Pragmatics
7. Multilinguality

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

- lectures: Murat Apishev, Denis Kirjanov, Katya Artemova
- seminars: Sergey Aksenov, Elizaveta Goncharova, Maria Sheyanova
- TAs: Anton Kukulyansky, Taisiya Glushkova, Dmitry Popov
- **Repo**: `github.com/PragmaticsLab/NLP-course-AMI`
- **Chat**: `t.me/nlp_cs`
- **Final mark**:
  $M_{1,2} = round(0.2quiz + 0.5HW + 0.3project)$
  $final = round(0.4exam + 0.3(M_1 + M_2))$
- **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

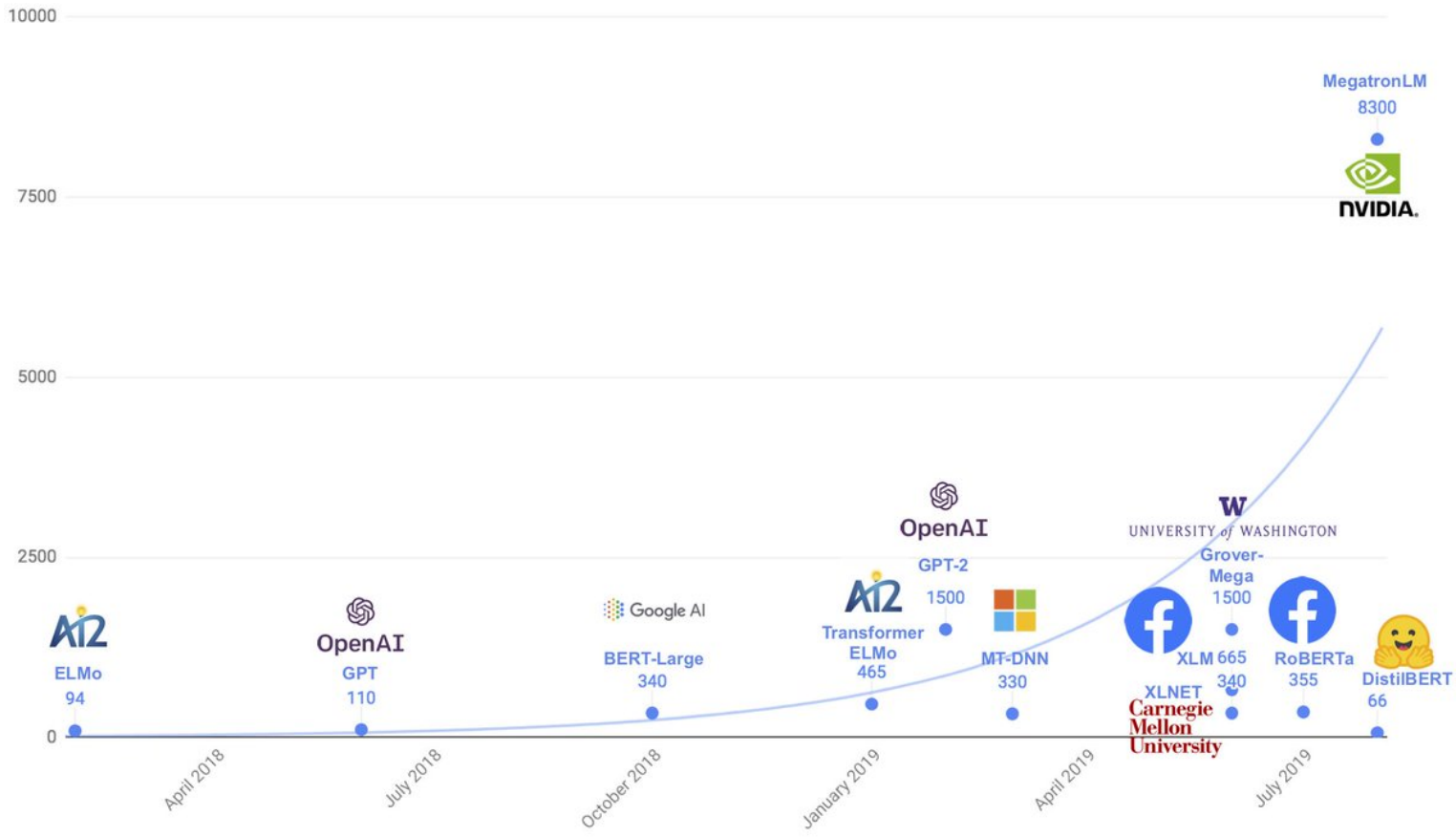
About this course

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]

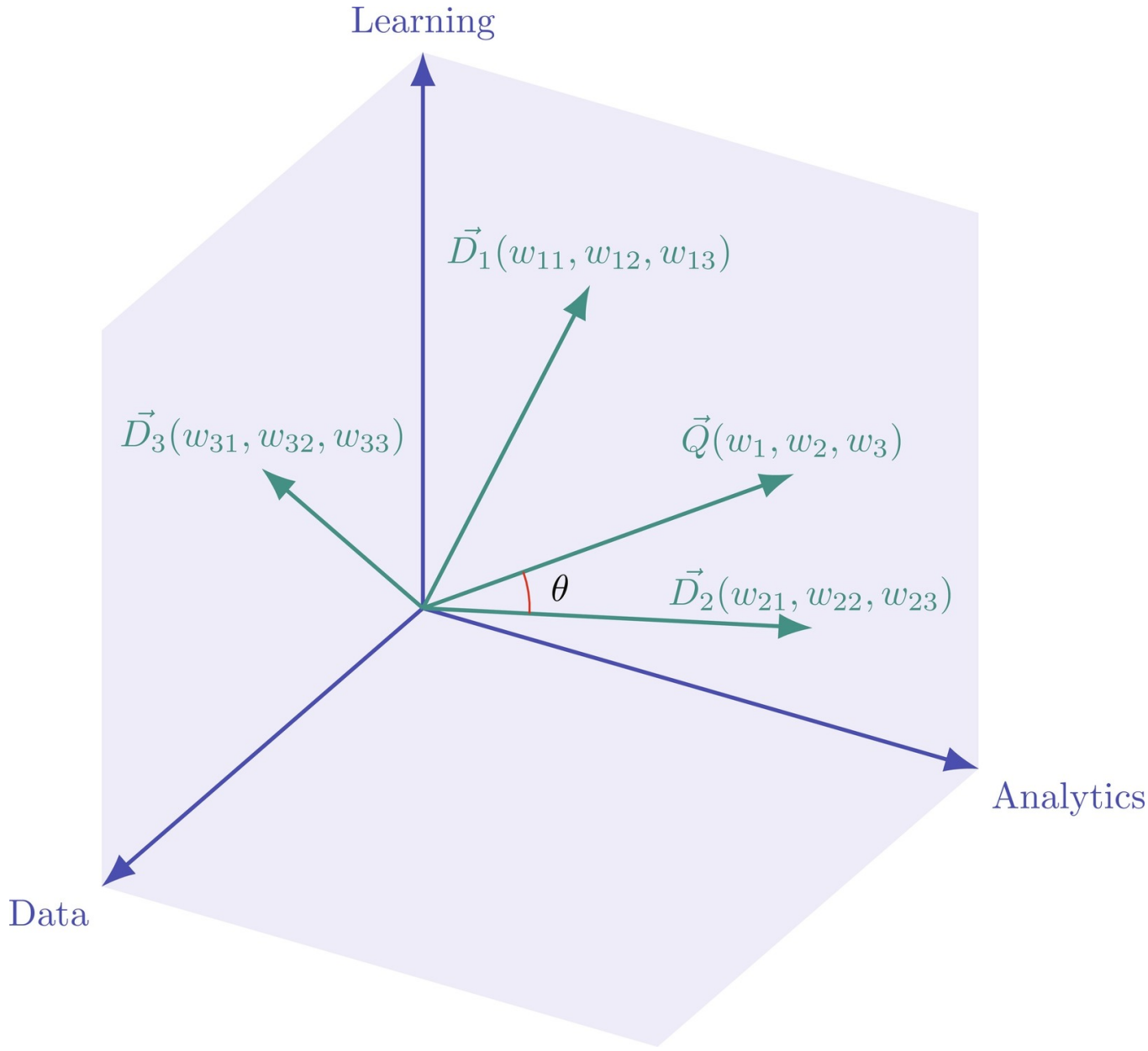


Image source: Handbook of statistics

Feed forward network

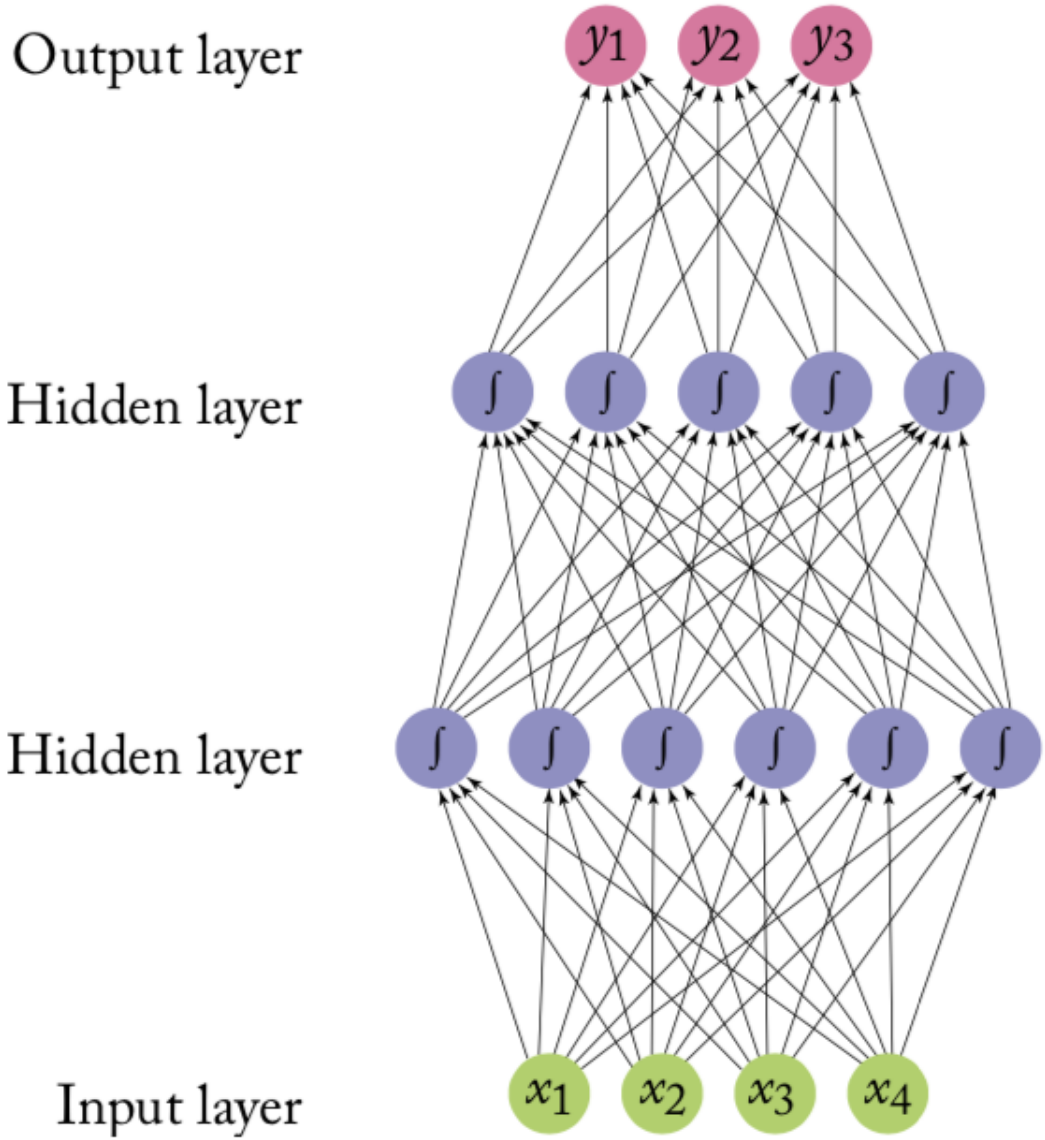
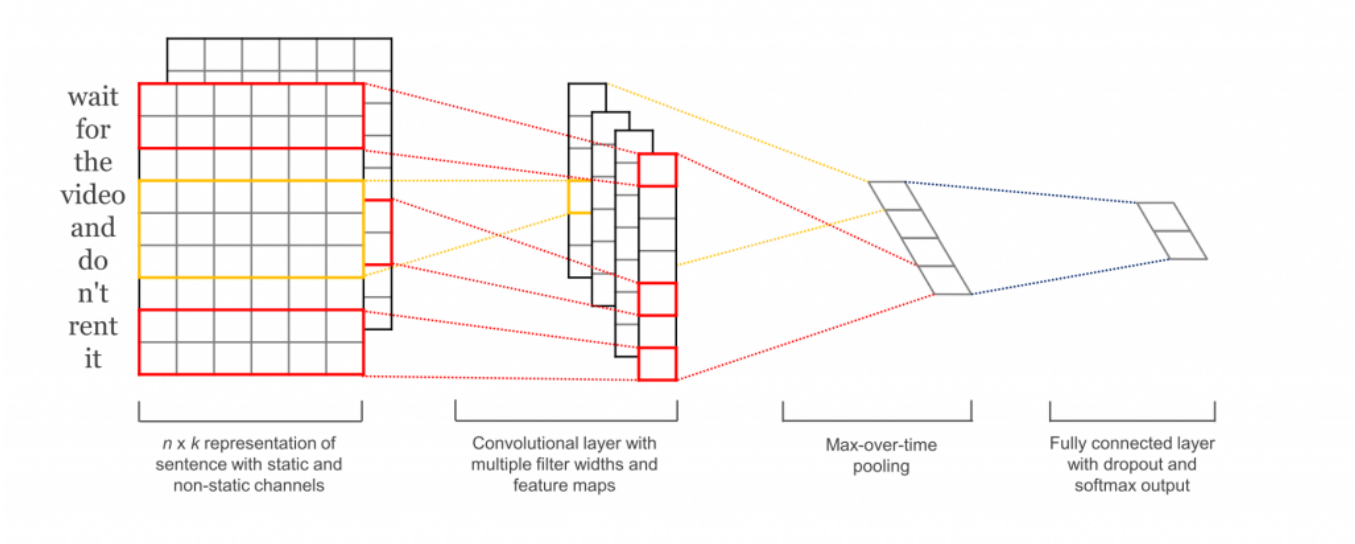


Image source: NN Methods for NLP

Convolutional network [2]



LSTM

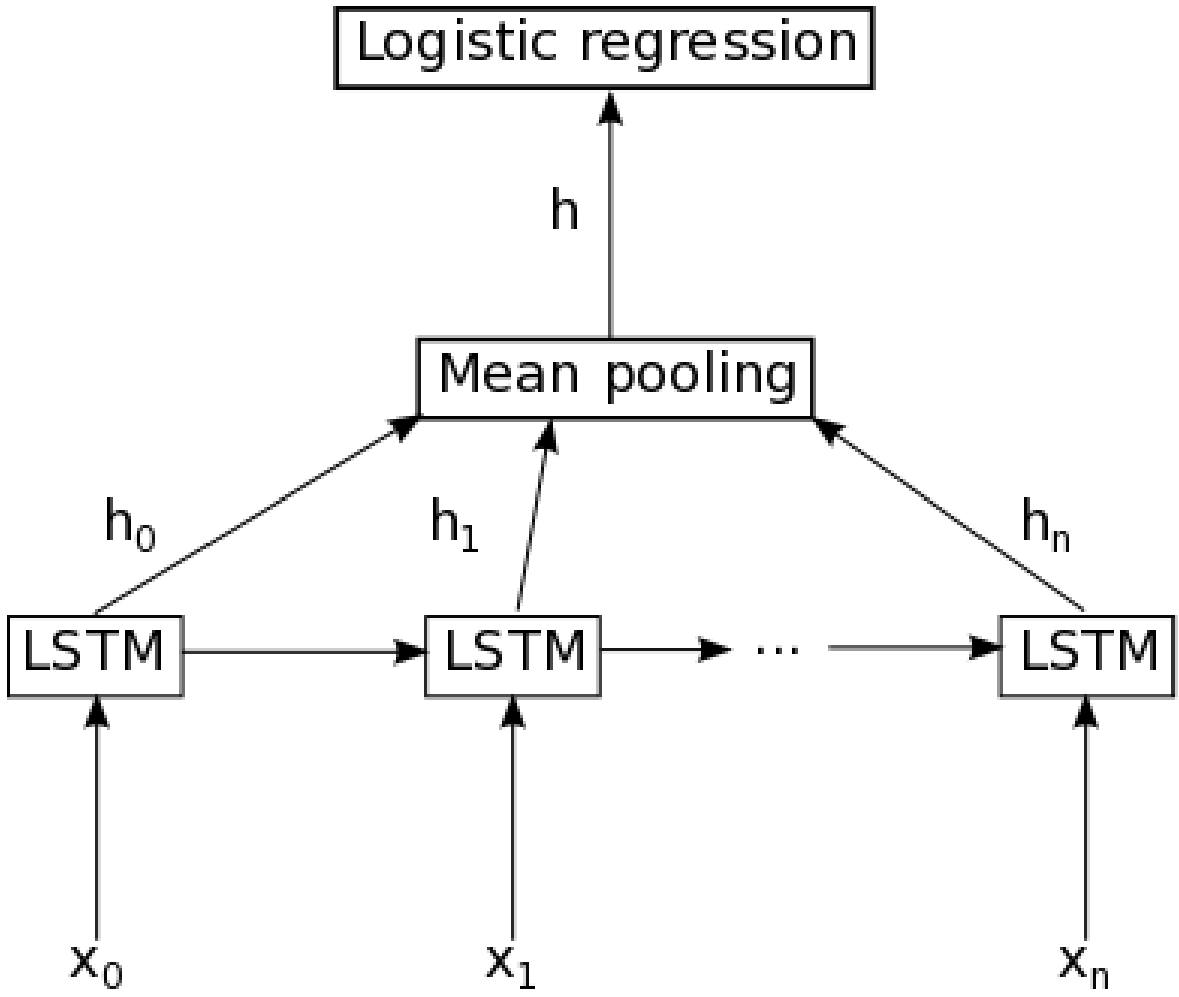
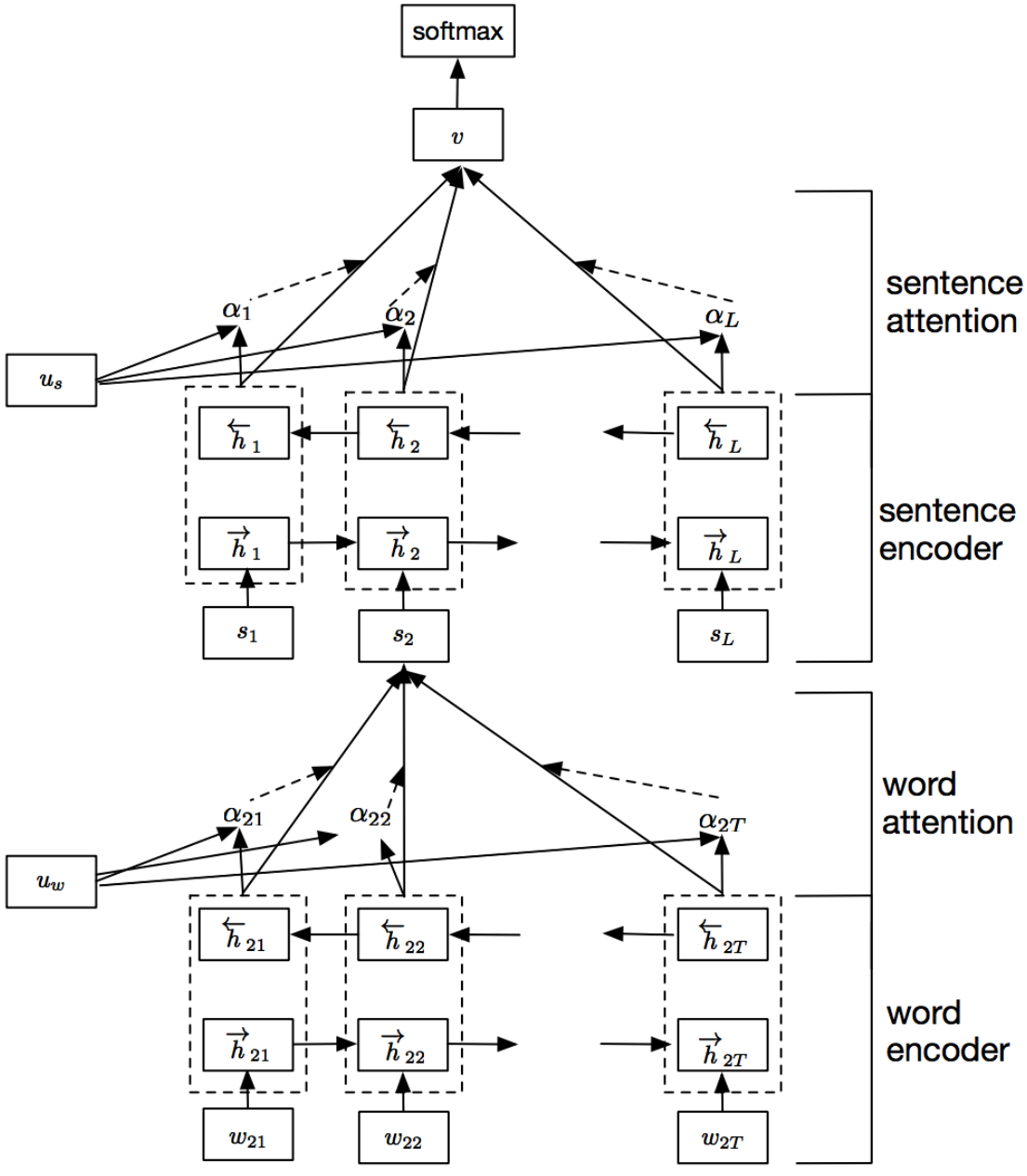
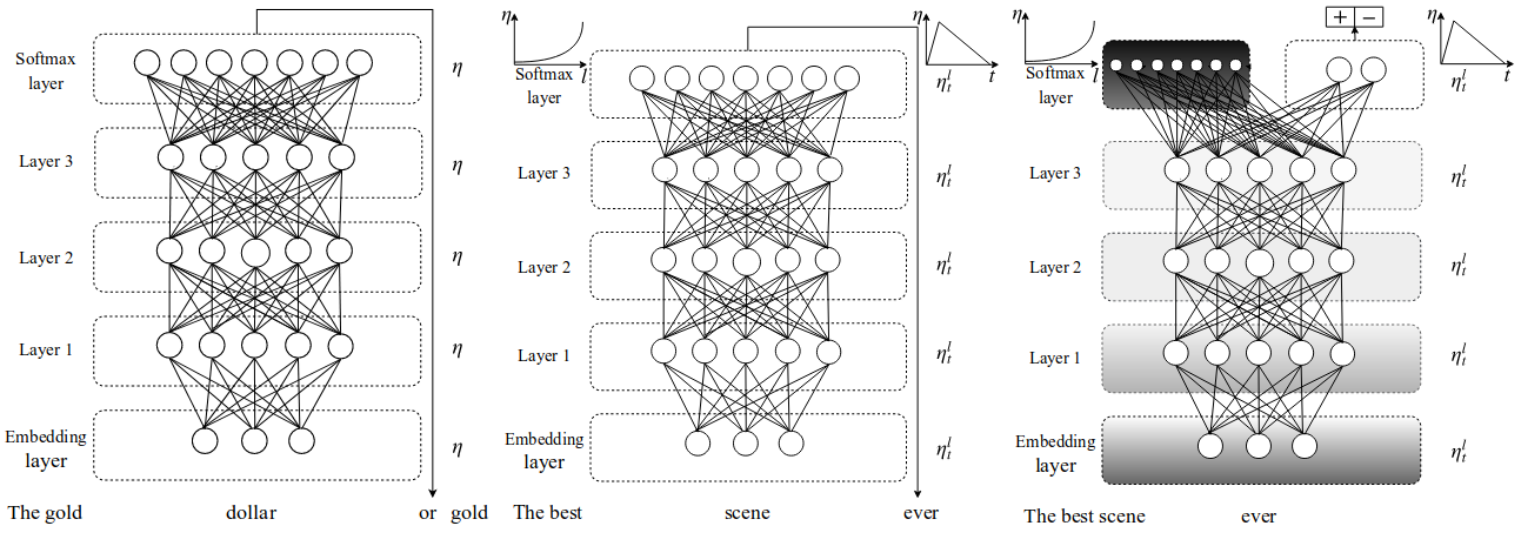


Image source: deeplearning.net

Hierarchical attention network [3]

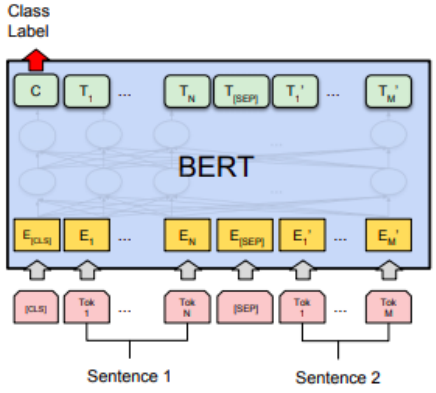


ULMFiT [4]

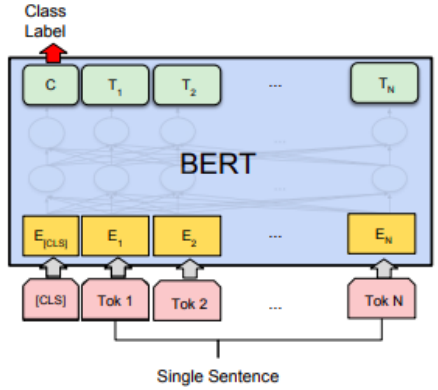


(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

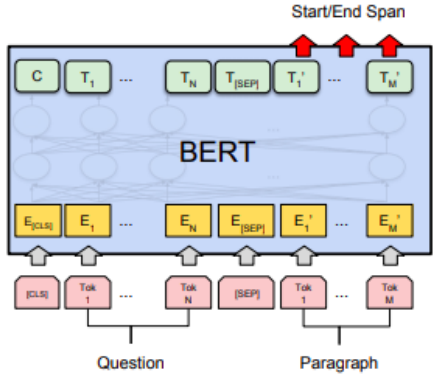
BERT [5]



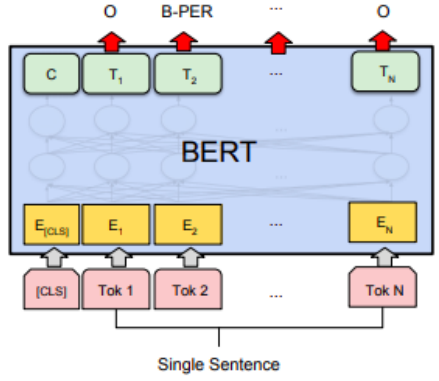
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(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.