Intro to NLP

Sergey Aksenov

Higher School of Economics

6 сентября 2021 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

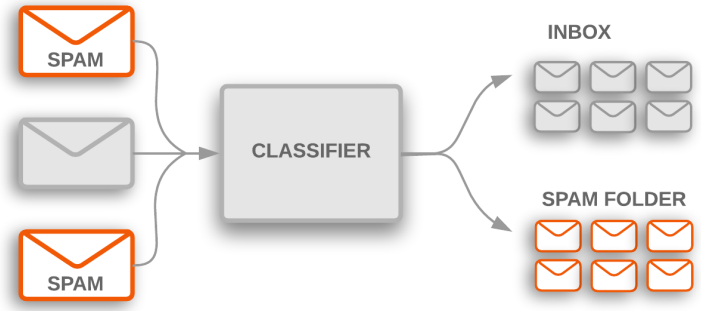
- Sentiment analysis
- Intent detection
- Spam filtering
- Topic classification

Sequence labelling

- Named entity recognition
- Coreference resolution

Sequence transformation (seq2seq)

- Machine translation
- Question answering



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

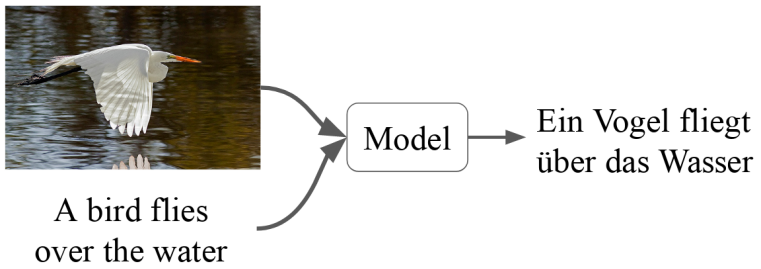


Image sources: developers.google.com, towardsdatascience.com, statmt.org

Phenomena to handle

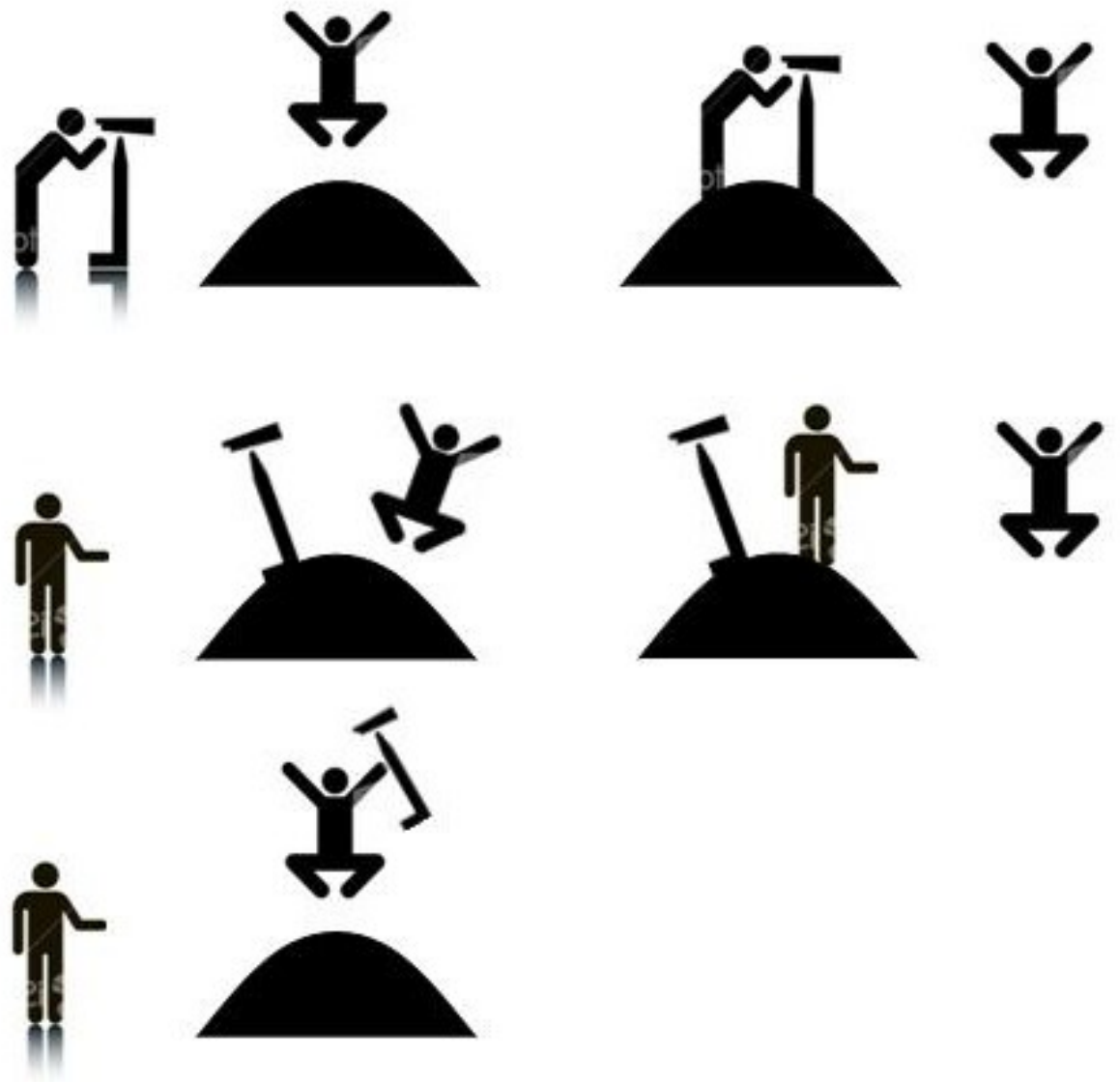
1. Tokenization and sentence boundary detection
2. Morphology
3. Syntax
4. Semantics
5. Discourse
6. Pragmatics
7. Multilinguality

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

- Teachers: Sergey Aksenov, Alena Fenogenova
- **Repo**:
  `https://github.com/Alenush/NLP-HSE_FinTechDA-2021`
- **Chat**: `https://t.me/joinchat/9Wds7C_sB19hYzE6`
- **Final mark**:
  $M_{hw} = \frac{1}{3}(M^1_{hw} + M^2_{hw} + M^3_{hw})$
  $M_{quiz} = \frac{1}{3}(M^1_{quiz} + M^2_{quiz} + M^3_{quiz})$
  $M_{final} = round(0.4M_{exam} + 0.7M_{hw} + 0.2M_{quiz})$
  $M_{exam}, M^i_{hw}, M^i_{quiz} \in [0, ..10]$

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

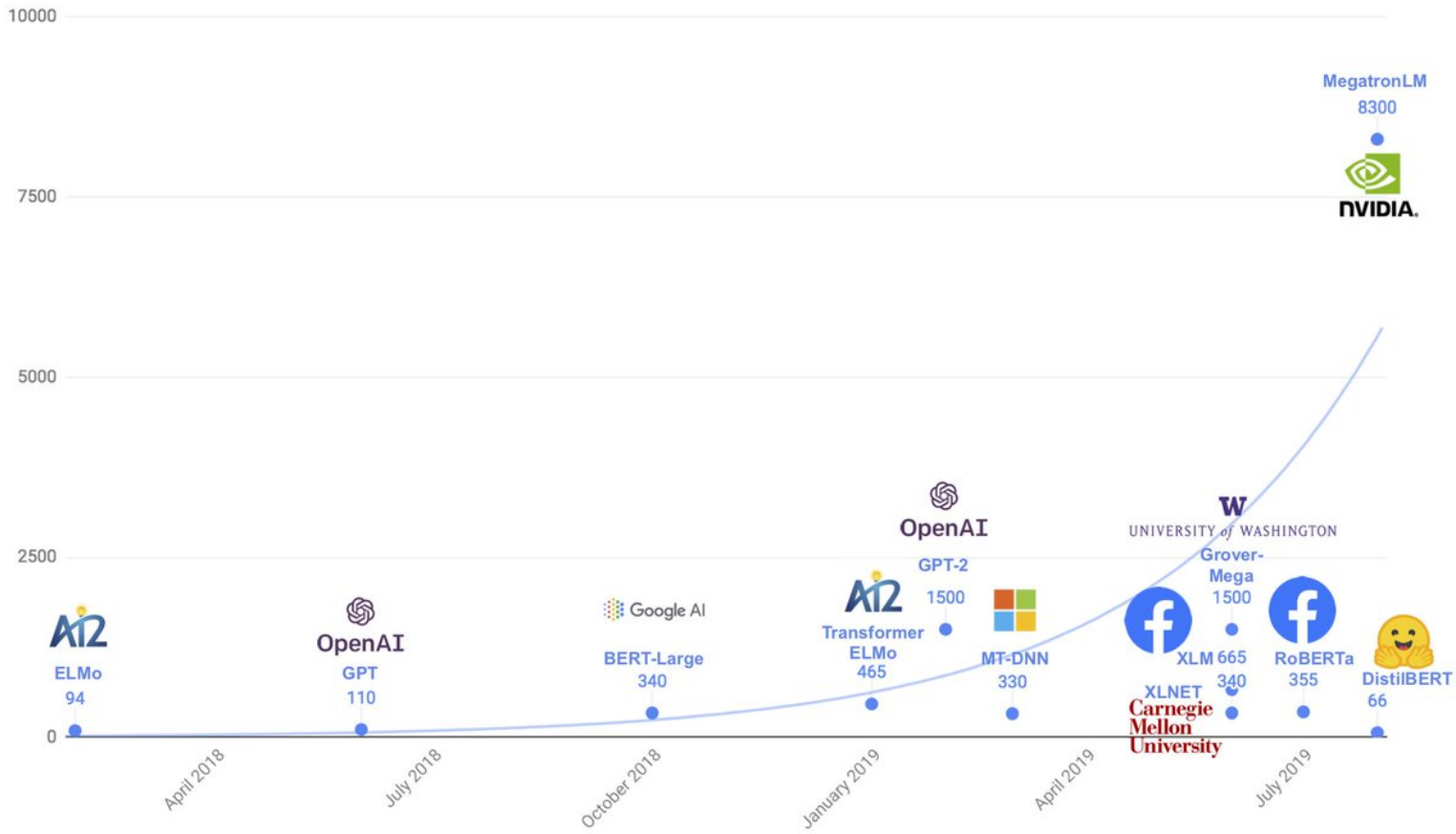
About this course

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. **Multilingual NLP**
6. **Programming Language Processing**
7. **Clinical NLP**

Today

Intro

About this course

Recent trends in NLP

Example task: text classification

Practice: tools for processing Russian

Vector space model **salton1975vector**

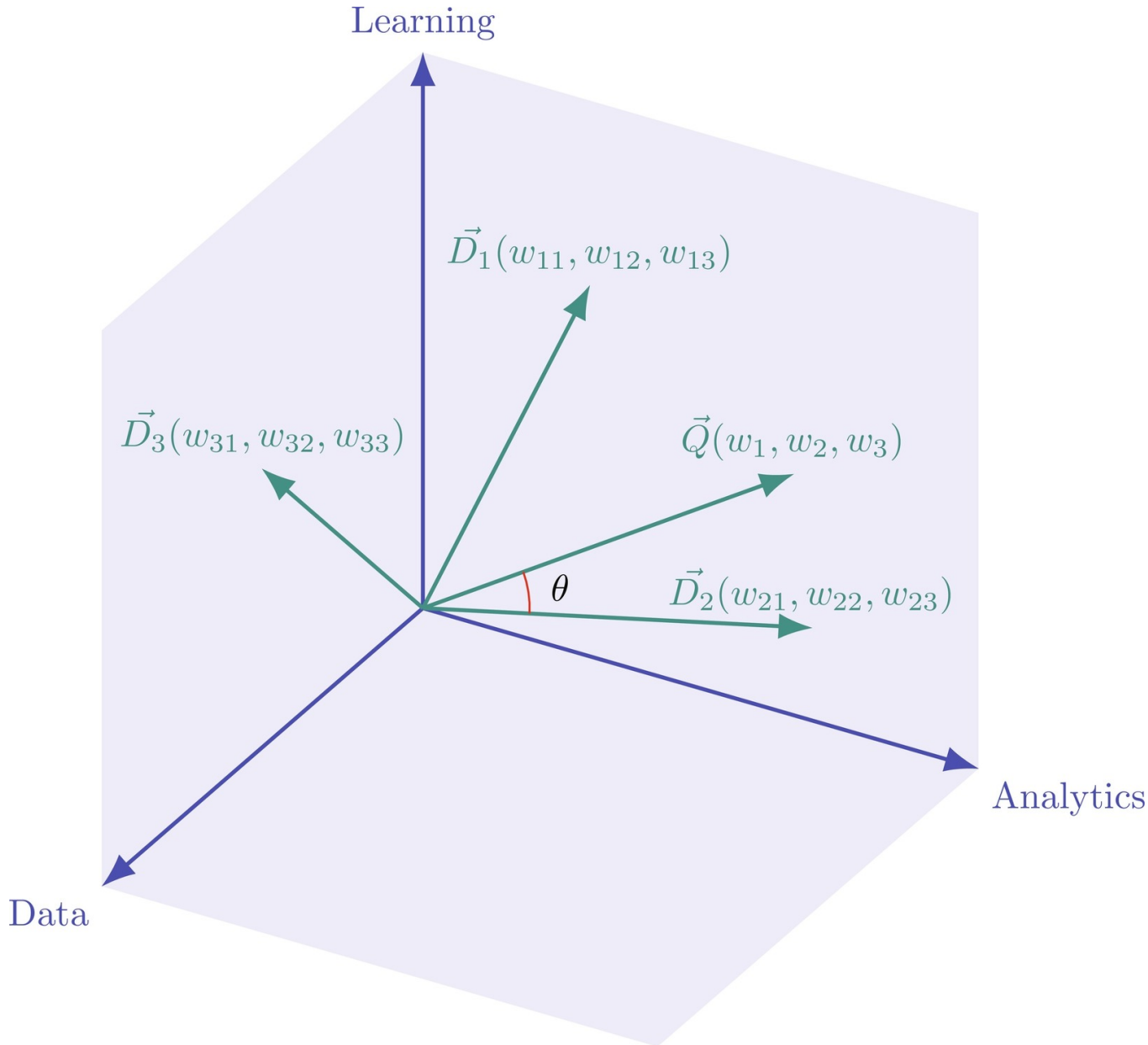


Image source: Handbook of statistics

Feed forward network

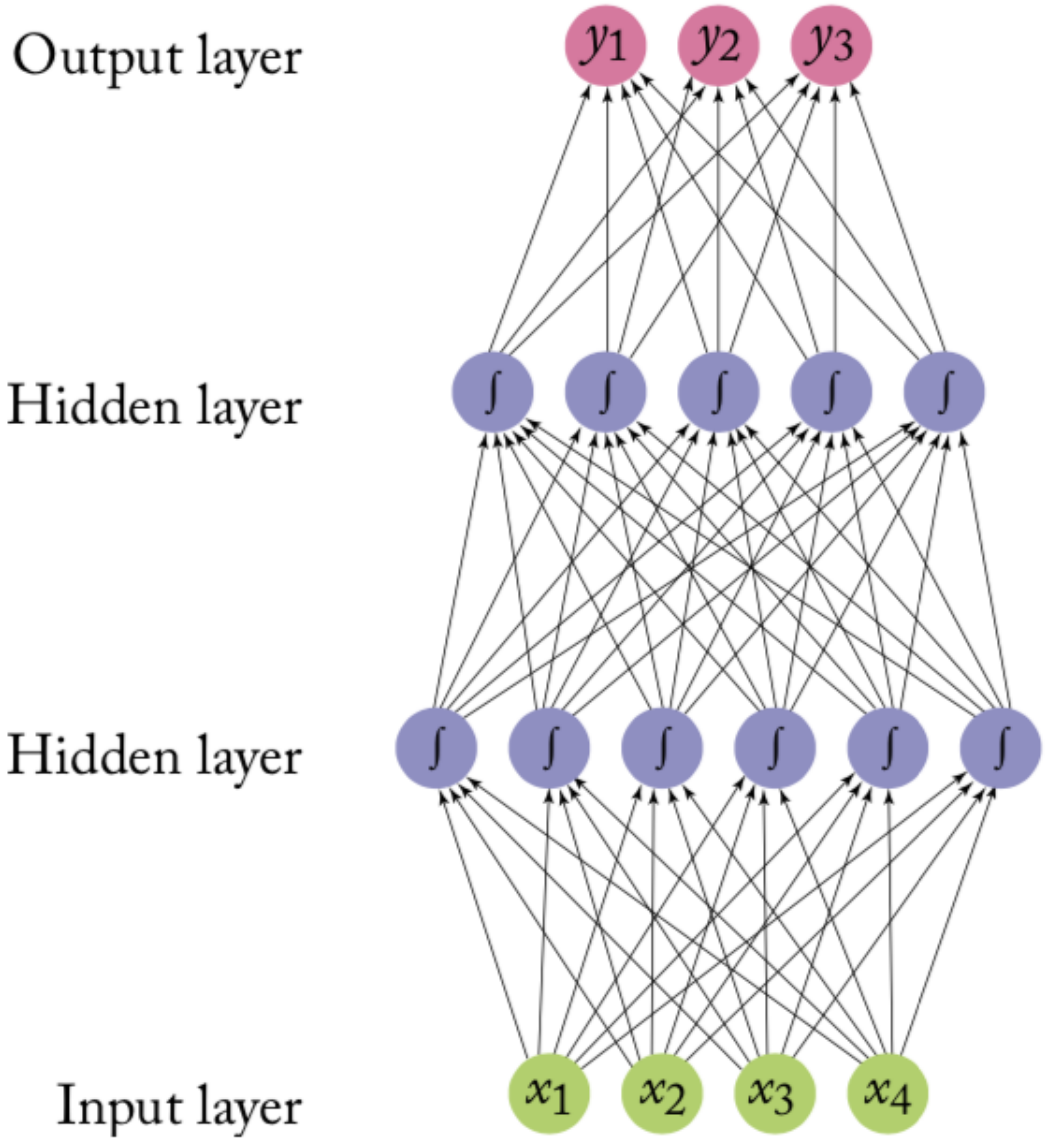
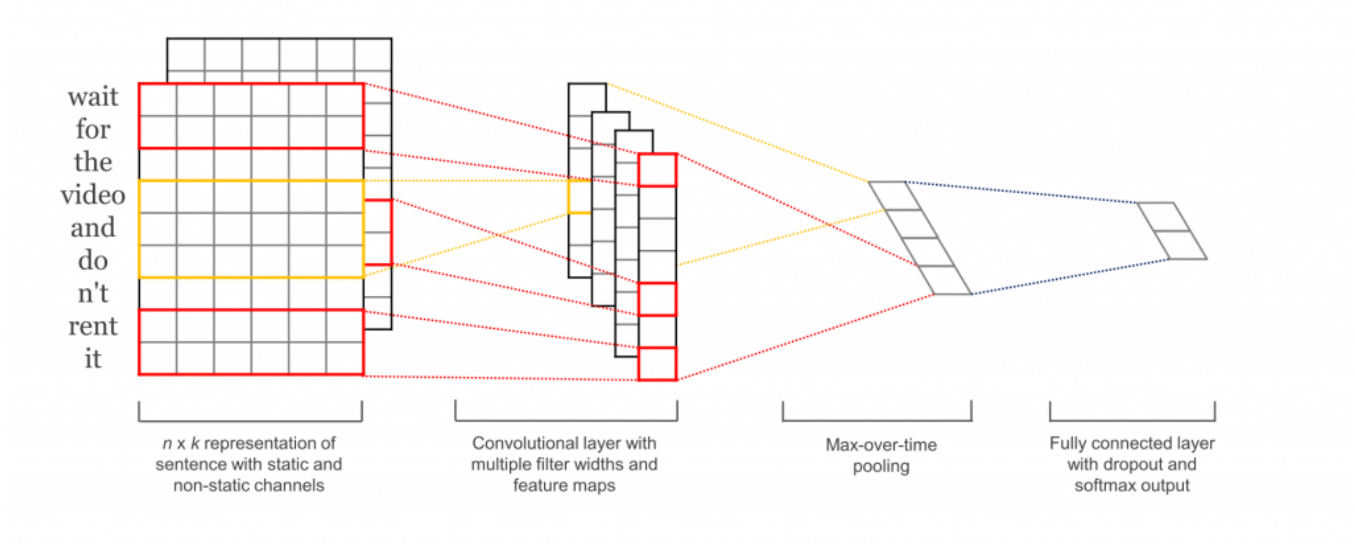


Image source: NN Methods for NLP

Convolutional network **kim2014convolutional**



LSTM

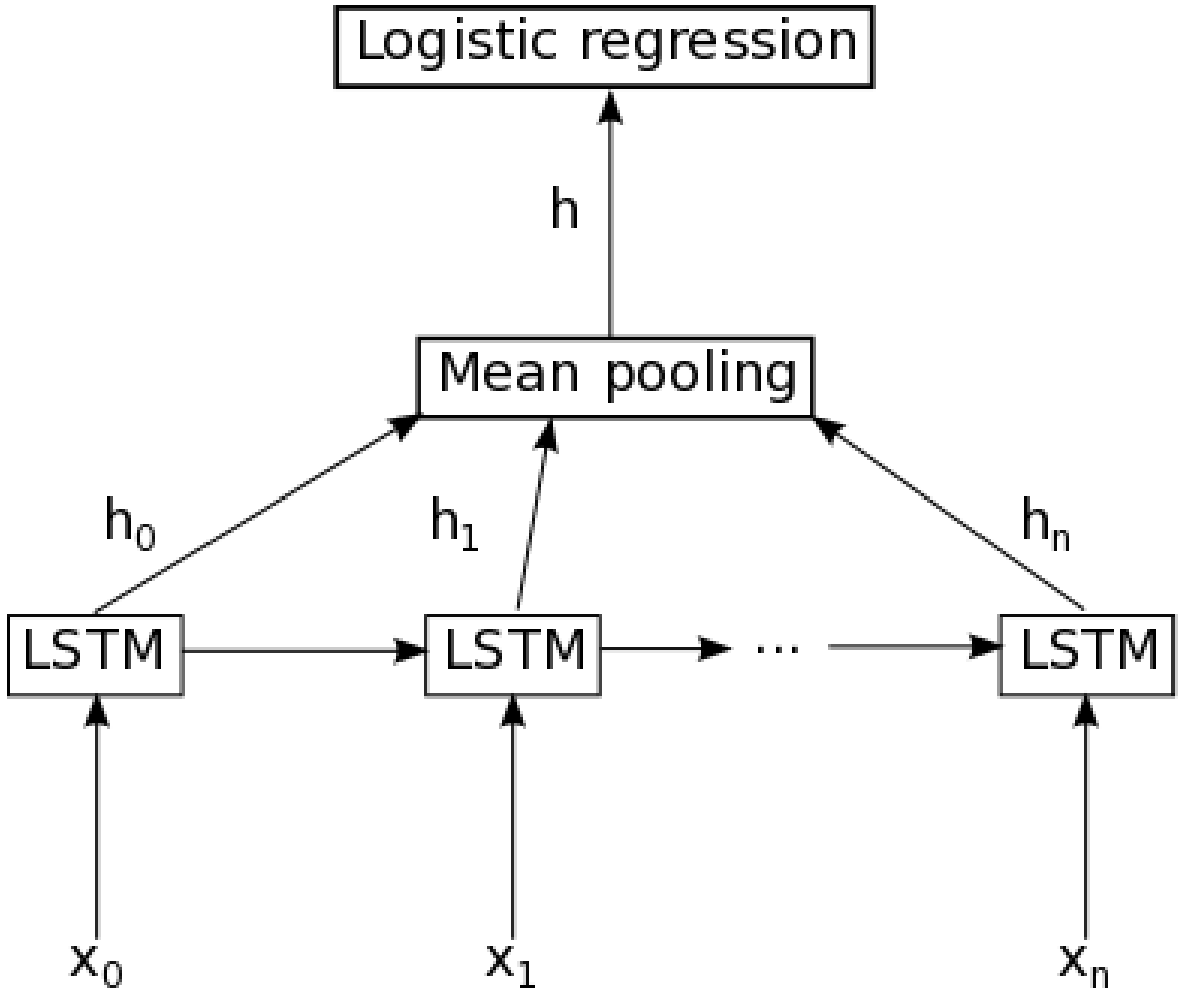
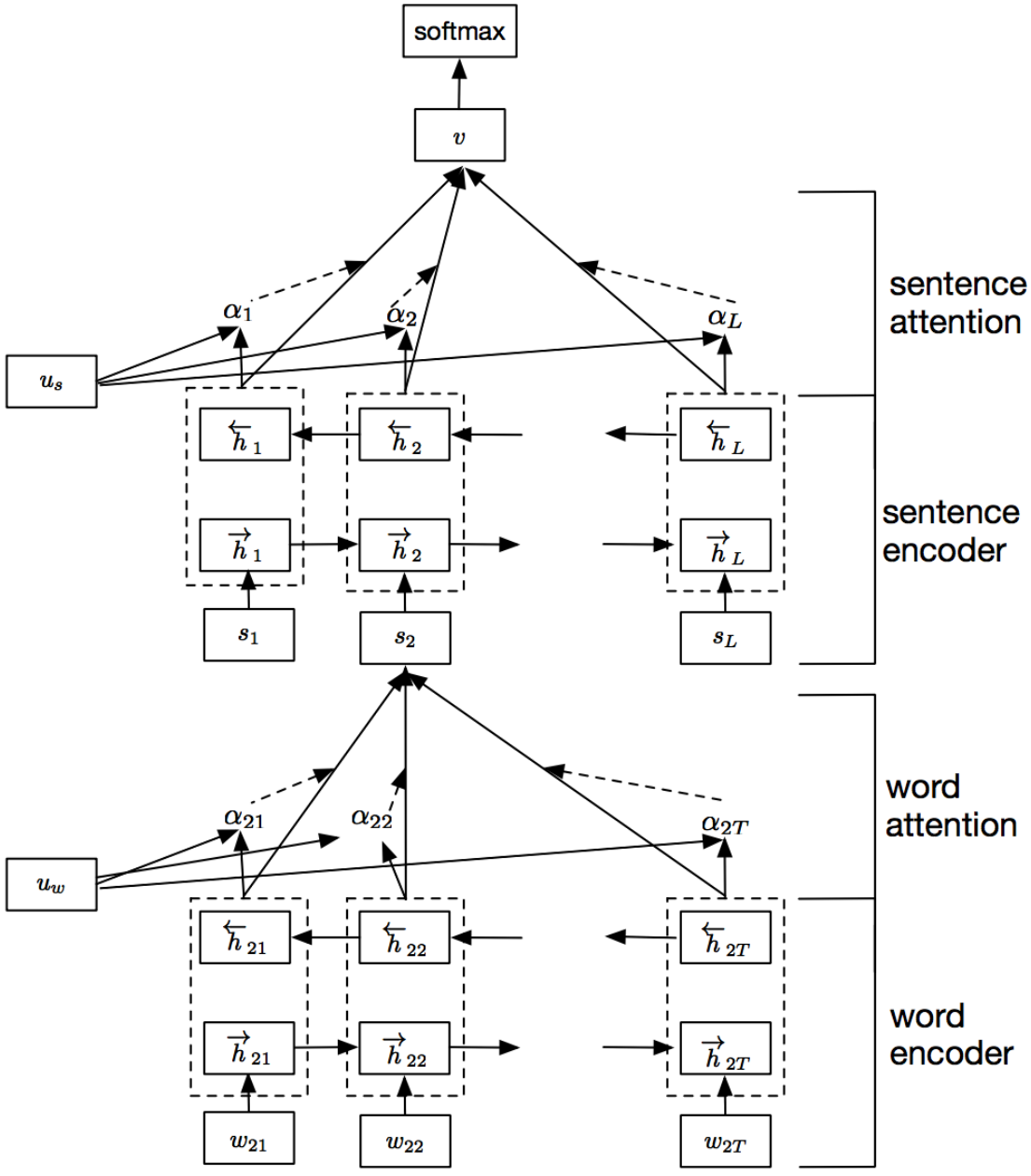
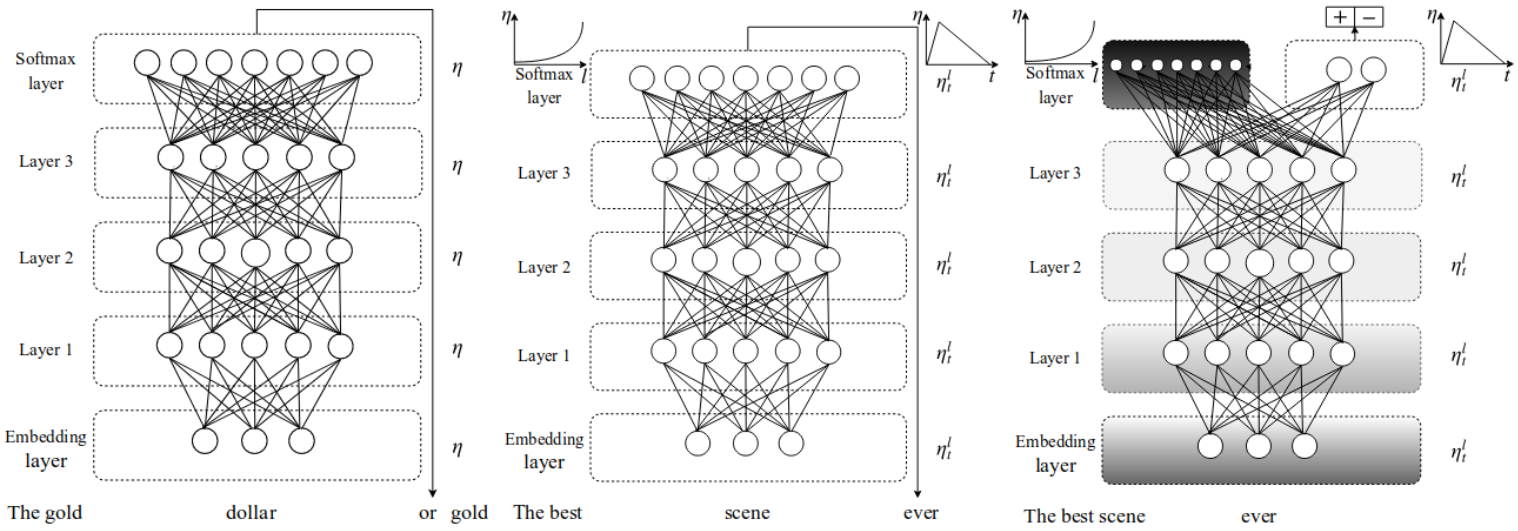


Image source: deeplearning.net

Hierarchical attention network **yang2016hierarchical**

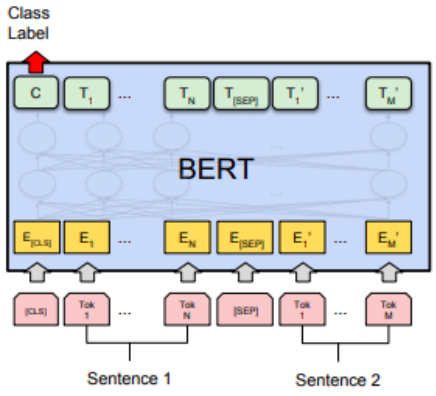


ULMFiT **howard2018universal**

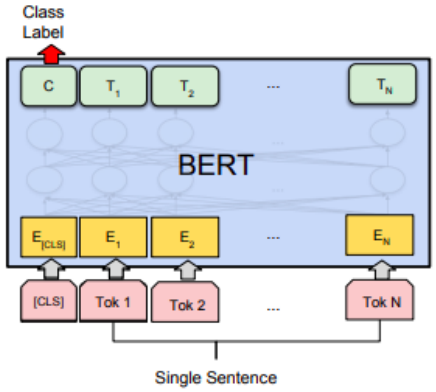


(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

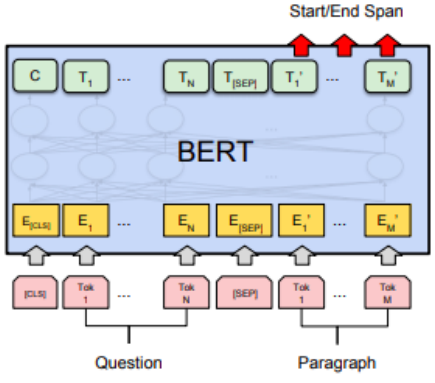
BERT **devlin2018bert**



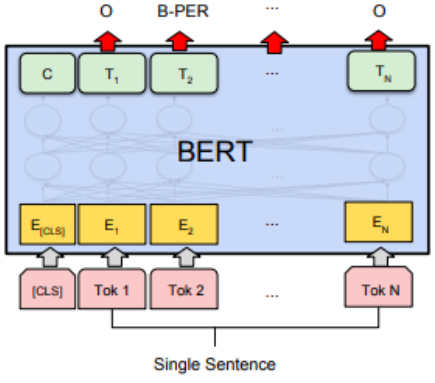
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



(c) Question Answering Tasks: SQuAD v1.1



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

OpenAI Codex

```
6  def decompress(compressed: str, tree: Tree) -> str:
7      '''
8      if compressed is empty return an empty string;
9      Create an empty string called result;
10     Index the dictionary tree with the first character of
   compressed; if k is a character append it to result; if k is
   a dictionary recursively call the compress function
11
12     '''
13     if compressed == "":
14         return ""
15     result = ""
16     tree_branch = tree[compressed[0]]
17     print(tree_branch)
18     for n in compressed[1:]:
19         if type(tree_branch) == str:
20             result += tree_branch
21             tree_branch = tree[n]
22         else:
23             tree_branch = tree_branch[n]
24     if type(tree_branch) == str:
25         result += tree_branch
26     return result
```

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