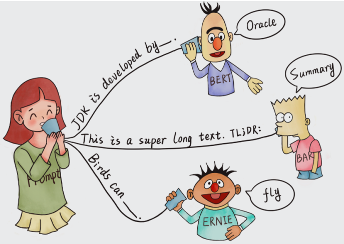
ML Basics

NLP tasks



Learning goals

- Understand the different types of tasks
- Linguistic tasks vs. higher-level tasks

CATEGORIZATION OF NLP TASKS

"Low-Level" tasks:

- *Token-level Classification*: Problems on a word/token level
- Modeling relationships between words/tokens (can also be rephrased to a token-level classification task)
- *Note:* The latter one can also be formulated in such a way, that it can be solved as a *token-level classification* task

"High-Level" tasks:

- *Sequence-level Classification*: Problems on a sequence level
- Producing sequences of text based on an input sequences, known as *seq2seq* tasks

EXAMPLES OF LOW-LEVEL NLP TASKS

Sequence tagging

- POS-tagging (part of speech)

Example

Time flies like an arrow.

Fruit flies like a banana.

EXAMPLES OF LOW-LEVEL NLP TASKS

Sequence tagging

- POS-tagging (part of speech)

Example

Time$_{\texttt{NN}}$ flies$_{\texttt{VBZ}}$ like$_{\texttt{IN}}$ an$_{\texttt{DT}}$ arrow$_{\texttt{NN}}$.
Fruit$_{\texttt{NN}}$ flies$_{\texttt{NN}}$ like$_{\texttt{VB}}$ a$_{\texttt{DT}}$ banana$_{\texttt{NN}}$.

IN = Preposition or subordinating conjunction (conjunction here); VBZ = Verb, 3rd person singular present; DT = determiner; NN = singular noun

LOW-LEVEL NLP TASKS

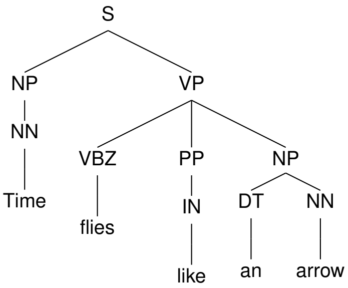
Sequence tagging

- POS-tagging (part of speech)

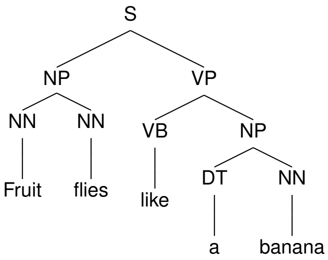
Structure prediction

- Chunking/Parsing

Example



Example



EXAMPLES OF LOW-LEVEL NLP TASKS

Sequence tagging

- POS-tagging (part of speech)

Structure prediction

- Chunking/Parsing

Semantics

- Word sense disambiguation

Example

Time flies like an arrow.

Fruit flies like a banana.

EXAMPLE: NAMED ENTITY RECOGNITION (NER)

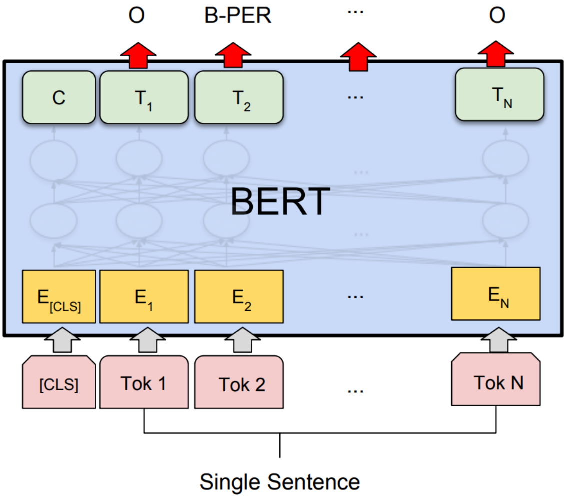
Example

"... chancelor$_{\mathtt{O}}$ Angela$_{\mathtt{B-PER}}$ Merkel$_{\mathtt{I-PER}}$ said$_{\mathtt{O}}$ ..."

"BIO"-tagging

- `B` = Begin of entity, e.g., `B-PER` (person), `B-LOC` (location)
- `I` = "Inside" entity, e.g, `I-PER`, `I-LOC`
- `O` = Other (no entity)

TOKEN-LEVEL CLASSIFICATION

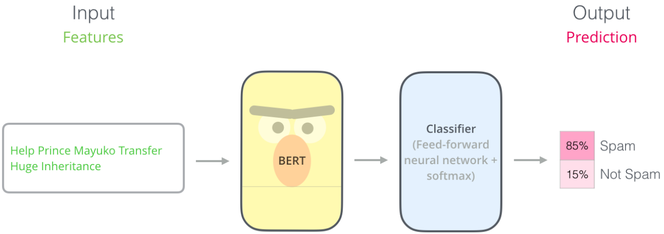


Source: *Devlin et al. (2018)*

HIGH-LEVEL NLP TASKS

- **Information Extraction**
  - search, event detection, textual entailment
- **Writing Assistance**
  - spell checking, grammar checking, auto-completion
- **Text Classification**
  - spam, sentiment, author, plagiarism
- **Natural language understanding**
  - metaphor analysis, argumentation mining, question-answering
- **Natural language generation**
  - summarization, tutoring systems, chat bots
- **Multilinguality**
  - machine translation, cross-lingual information retrieval

SEQUENCE-LEVEL CLASSIFICATION



Source: *Jay Alammar*

Notes:

- BERT is a popular model, no need to know further details now
- Output can also be non-binary, i.e. multi-class/-label

EXAMPLE: MACHINE TRANSLATION

A brief History of Machine Translation

- Rule-Based Machine Translation (50s – 80s)
  - Dictionaries + Grammatical Rules
- Example-Based Machine Translation (80s – 90s)
  - First suggested by Makoto Nagao (1984)
  - Based on bilingual text corpora
- Statistical Machine Translation (90s – 10s)
  - Driven by IBM guys
- Neural Machine Translation (last few years)
  - Based on neural networks (LSTMs, Transformers)

SEQ2SEQ MODELING

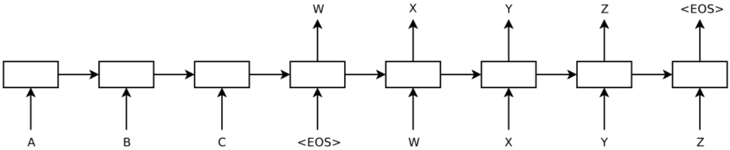


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Source: *Sutskever et al. (2014)*

Notes:

- In the meantime: Transformers replaced LSTMs
- Overall architecture (*Encoder-Decoder*) still used

SEQ2SEQ MODELING

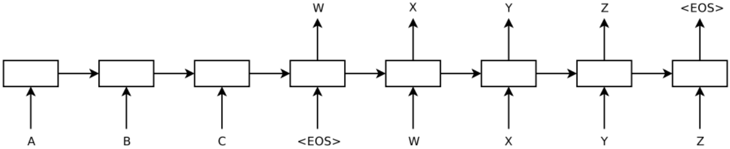


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Source: *Sutskever et al. (2014)*

Used for:

- (Neural) Machine Translation
- Summarization
- Questions answering

BENCHMARKING: GLUE

- Nine sentence- or sentence-pair language understanding tasks
- Public leaderboard, (still) very popular benchmark collection

Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

Source: *Wang et al. (2018)*

BENCHMARKING: GLUE

- **CoLa**
  - Short for "Corpus of Linguistic Acceptability"
  - Binary classification: Linguistically acceptable or not
- **SST-2 (Stanford Sentiment Treebank)**
  - Movie reviews annotated with their sentiment (pos/neg)
  - Also available in a more fine-grained fashion (SST-5)
- **MRPC (Microsoft Research Paraphrase Corpus)**
  - Sentence pairs; Classify whether the are semantically equivalent
- **QQP (Quora Question Pairs)**
  - Sentence pairs; Classify whether the are semantically equivalent
  - Here: Questions (as opposed to "news" in MRPC)

BENCHMARKING: GLUE

- **STS-B (Semantic Textual Similarity Benchmark)**
  - Regression task (similarity score from 1 to 5)
- **MNLI (Multi-Genre Natural Language Inference)**
  - Textual Entailment: Premise and Hypothesis are given
  - Target: *entailment / contradiction / neutral*
- **QNLI (Stanford Question Answering Dataset)**
  - Modified original task to binary classification
- **RTE (Recognizing Textual Entailment)**
  - Collapsed to a binary classification task
- **WNLI (Winograd Schema Challenge)**
  - Coreference resolution task
  - Rephrased to entailment task by providing the model with original sentence and sentence with the pronoun