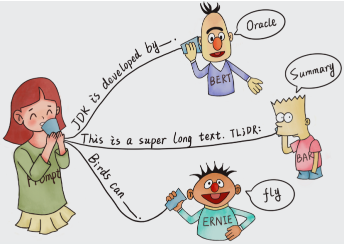
Basics

NLP tasks



Learning goals

- Understand the different types of tasks (low- vs. high-level)
- Purely Linguistic tasks vs. more general classification tasks

CATEGORIZATION OF NLP TASKS

Distinction between:

- Language modeling
- Token-level classification
- Sequence-level classification
- Similarity / Retrieval
- Text generation

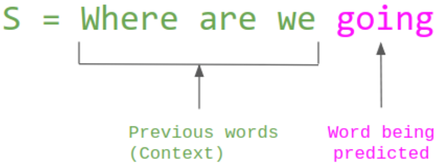
Connection to learning paradigms:

- Given the task, some learning paradigms are more suitable
- Tasks can be formulated differently to fit a given learning paradigm
- Amount of available (labeled) data might depend on task
- Presence/Absence of labels important to consider

LANGUAGE MODELING

Predict the next token:

$$
\begin{aligned}
P_{(w_1, w_2, \ldots, w_n)} &= p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)...p(w_n|w_1, w_2, .., w_{n-1}) \\
&= \prod_{i=1}^{n} p(w_i|w_1, ..., w_{i-1})
\end{aligned}
\tag{1}
$$



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Source: The Gradient

CATEGORIZATION OF NLP TASKS

"Low-Level" tasks:

- *Token-level Classification*: Problems on a word/token level
- Modeling relationships *between* words/tokens
- *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
- *Retrieval*: Assess (semantic) similarity on document-level
- Producing sequences of text based on an input sequences, known as *seq2seq* tasks
- *Note:* The latter one is also an instance of a generation task

LOW-LEVEL: SEQUENCE TAGGING

POS-tagging (part of speech):

Example

Time flies like an arrow.
Fruit flies like a banana.

LOW-LEVEL: SEQUENCE TAGGING

POS-tagging (part of speech):

Example

Time flies like an arrow.
Fruit flies like a banana.

Example

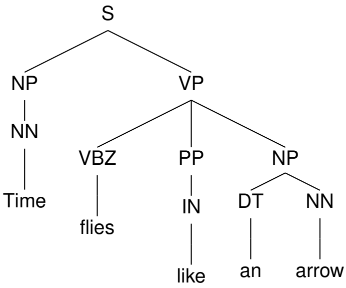
Time$_{\texttt{NN}}$ flies$_{\texttt{VBZ}}$ like$_{\texttt{IN}}$ an$_{\texttt{DT}}$ arrow$_{\texttt{NN}}$.
Fruit$_{\texttt{NN}}$ flies$_{\texttt{NN}}$ like$_{\texttt{VB}}$ a$_{\texttt{DT}}$ banana$_{\texttt{NN}}$.

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

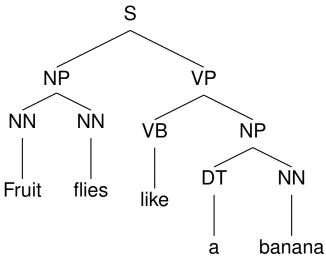
LOW-LEVEL: STRUCTURE PREDICTION

Chunking/Parsing:

Example



Example



LOW-LEVEL: SEMANTICS

Word sense disambiguation:

Example

Time flies like an arrow.

Fruit flies like a banana.

NAMED ENTITY RECOGNITION (NER)

Example

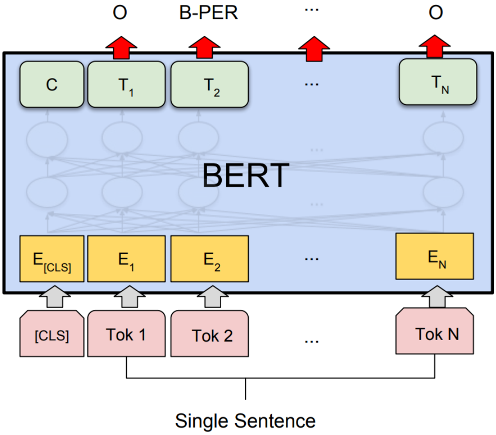
"... chancelor$_{\mathtt{O}}$ Angela$_{\mathtt{B-PER}}$ Merkel$_{\mathtt{I-PER}}$ said$_{\mathtt{O}}$ ..."

"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

Pre-train/fine-tune:

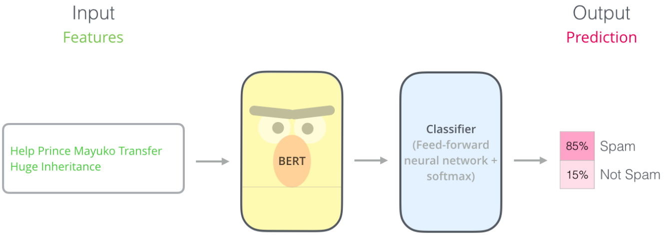


▸ Source: Devlin et al., 2019

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



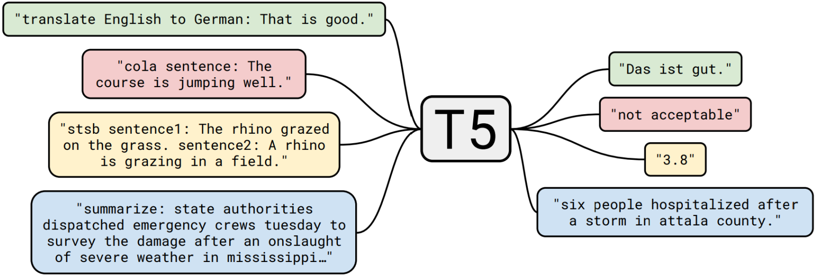
Source: Jay Alammar

Notes:

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

SEQUENCE-LEVEL CLASSIFICATION

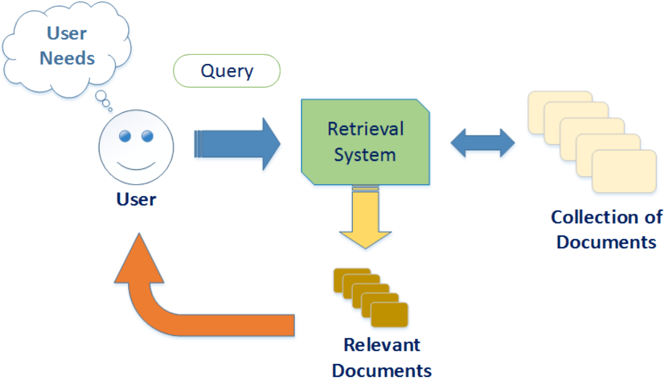
Reformulation as generative task:



▸ Source: Raffel et al., 2020

RETRIVAL (CF. PREVIOUS CHAPTER)

Document retrieval



▸ Source: Analytics Vidhya

GENERATION: 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 (10s – now)
  - Based on neural networks (LSTMs, Transformers)

SEQ2SEQ MODELING

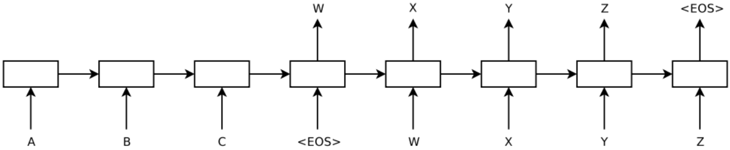


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that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

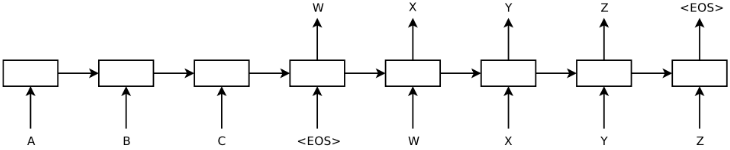


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that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

▸ Source: Sutskever et al., 2014

Used for:

- (Neural) Machine Translation
- Summarization
- Questions answering

BENCHMARKING: TRADITIONAL NLU

- 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: GENERATION

- **WinoGrande**

Dataset Viewer

Auto-converted to Parquet API

Subset: winogrande_debiased (12.3k rows)

Split: train (9.25k rows)

sentence (string)	option1 (string)	option2 (string)	answer (string)
"John moved the couch from the garage to the backyard to create space. The _ is small."	"garage"	"backyard"	"1"
"The doctor diagnosed Justin with bipolar and Robert with anxiety. _ had terrible nerves recently."	"Justin"	"Robert"	"2"
"Dennis drew up a business proposal to present to Logan because _ wants his investment."	"Dennis"	"Logan"	"1"
"Felicia unexpectedly made fried eggs for breakfast in the morning for Katrina and now _ owes a favor."	"Felicia"	"Katrina"	"2"

▸ Source: Sakaguchi et al., 2019

- **HellaSwag**

Pick the best ending to the context.

How to catch dragonflies. Use a long-handled aerial net with a wide opening. Select an aerial net that is 18 inches (46 cm) in diameter or larger. Look for one with a nice long handle.

a) Loop 1 piece of ribbon over the handle. Place the hose or hose on your net and tie the string securely.	b) Reach up into the net with your feet. Move your body and head forward when you lift up your feet.	c) If possible, choose a dark-colored net over a light one. Darker nets are more difficult for dragonflies to see, making the net more difficult to avoid.	d) If it's not strong enough for you to handle, use a hand held net with one end shorter than the other. The net should have holes in the bottom of the net.

▸ Source: Zellers et al., 2019

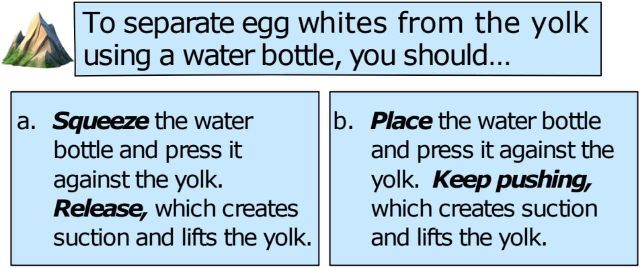
BENCHMARKING: GENERATION

- **LAMBADA**

(5) *Context:* He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I ’ve already focused on my friend. You just have to click the shutter, on top, here.”
Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

▸ Source: Paperno et al., 2016

- **PIQA**



▸ Source: Bisk et al., 2019