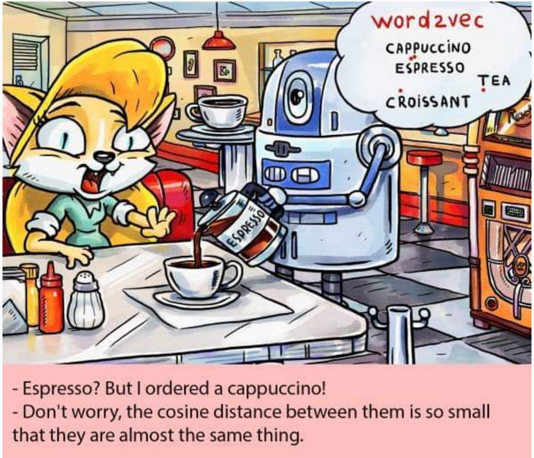
Modern NLP @ work

A (short) historical overview & recent Applications

Matthias Aßenmacher (LMU München)

11. März 2021



That's not how it's intended to work ..

Recent advances in NLP

We encounter more and more NLP applications in everyday life:

- Chatbots are on the rise
- Alexa or Siri have become standard tools
- GoogleTranslate or DeepL are commonly used

The world's largest Tech companies are investing heavily:

- fb ai research, google ai, microsoft research have own NLP groups
- Leading researchers like Geoffrey Hinton (Google) or Yann LeCun (Facebook) start working for the industry

How to represent words? – The distributional hypothesis

Zellig S. Harris (1954):

Distributional Structure

J.R. Firth (1957):

"You shall know a word by the company it keeps."

Learn something about the meaning of *football* by studying which context it appears in:

.. the score of the	*football*	game was 3:0 ..
.. he shot the	*football*	directly at the goalkeeper ..
.. last night, I was watching	*football*	on tv ..

One-hot vs. context-based encoding

One-hot encoding:

$$football = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$
$$basketball = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

Two major problems:

- $similarity(football, basketball) = ?$
  The vectors are orthogonal to each other, so $sim(w_i, w_j) = 0 \; \forall \; i, j$
- The dimensionality of these vectors?

One-hot vs. context-based encoding

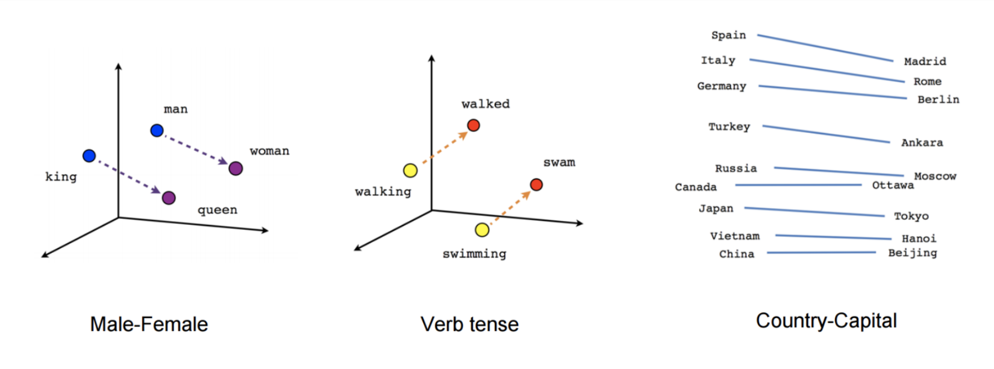
Context-based encoding:

$$\begin{aligned} \textit{football} &= [0, 3, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 2, 1, 4] \\ \textit{basketball} &= [0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 1, 0, 3, 3, 2] \end{aligned}$$

Two major problems:

- $\textit{similarity}(\textit{football}, \textit{basketball}) = ?$
- The dimensionality of these vectors?

The breakthrough: *Word embeddings*



Source: towardsdatascience

How did they do this?

The "Fake Task":

- *Training objective:* Given a word, predict the neighbouring words
- *Generation of samples:* Sliding fixed-size window over the text

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (the, quick); (the, brown)

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (quick, the); (quick, brown); (quick, fox)

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (brown, the); (brown, quick), (brown, fox), (brown, jumps)

A note one the literature

Four papers out there:

- About the model architectures:

  ▸ Mikolov et al. (2013a)
- Computational subtleties:

  ▸ Mikolov et al. (2013b)
- Linguistic Regularities:

  ▸ Mikolov et al. (2013c)
- Use for Machine Translation:

  ▸ Mikolov et al. (2013d)

Context-free Embeddings

1st Generation of neural embeddings are *"context-free"*

- Breakthrough paper by Mikolov et al, 2013 (Word2Vec)
- Followed by Pennington et al, 2014 (GloVe)
- Extension of Word2Vec by Bojanowski et al, 2016 (FastText)

Why "Context-free"?

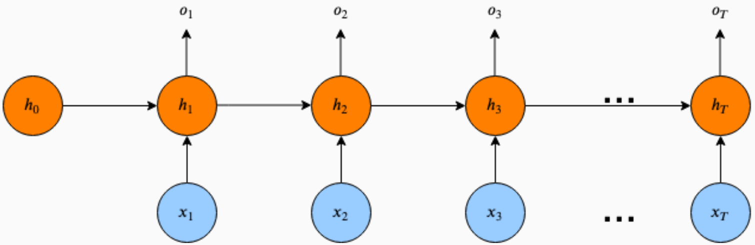
- Models learn *one single* embedding for each word
- Why could this possibly be problematic?
  - "The *default* setting of the function is xyz."
  - "The probability of *default* is rather high."
- Would be nice to have different embeddings for these two occurrences

Contextual Embeddings

“Contextual” embeddings?

- Model makes further use of the context a word appears in
- Embeddings depend on the context around a word
- Requires us to process sequences $\rightarrow$ **RNNs/LSTMs:**
  - Take sequences (e.g. time series) as inputs
  - But also: Sequences of characters, words or tokens
- Distinguish between *Uni- & Bidirectional* contextuality

Processing one part of the input at a time:



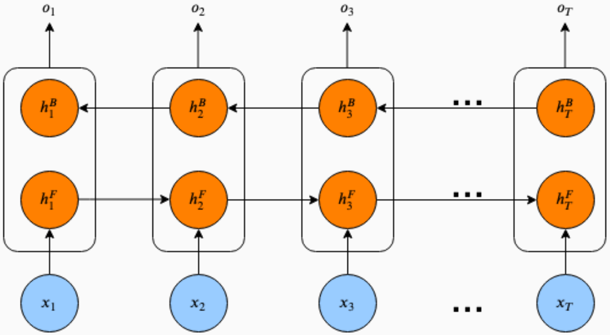
An unrolled (unidirectional) recurrent neural network

Bidirectionality

Why bidirectionality?

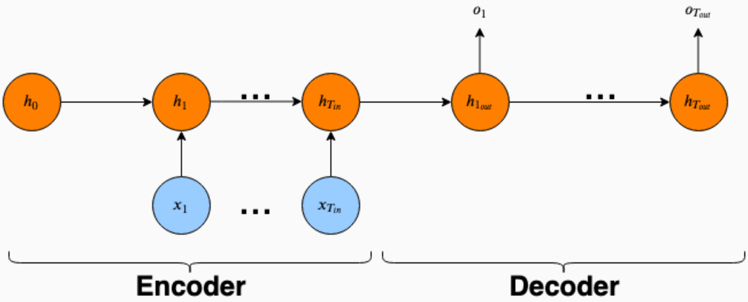
- Vanilla RNNs/LSTMs just capture the left hand context
- This might make sense when considering the language modelling objective
- *Counterexample:* Machine Translation
  $\rightarrow$ Translation of a word might also depend on the right hand context

Simultaneously running a backward RNN:



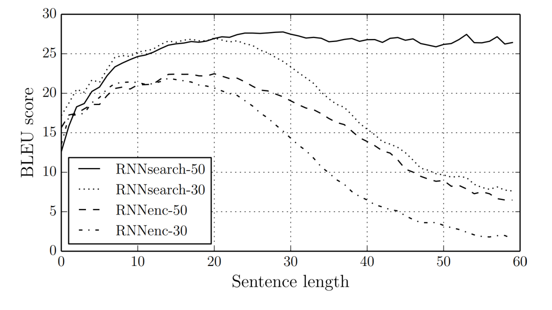
An unrolled bidirectional recurrent neural network

Graphical illustration:



An unrolled (unidirectional) encoder-decoder RNN

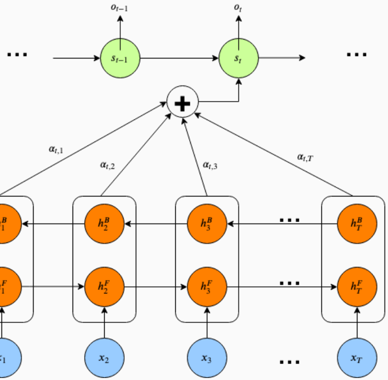
Long-range dependencies



Source: Bahdanau et al. (2014)

- RNNenc (classical encoder-decoder); RNNsearch (with Attention)
- BLEU score: Measure for translation quality (higher is better)

Bahdanau et al. (2014)



Use weighted combinations of all the (concatenated) hidden states

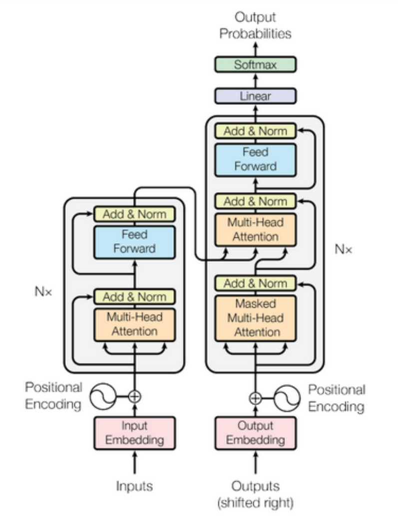
Attention Is All You Need

The basic ingredients:

- Model architecture introduced by ‣ Vaswani et al. (2017)
- Encoder-Decoder framework relying completely on Self-Attention
  $\rightarrow$ **No** recurrence at any place in the network
- Requires large matrix multiplications, **but:** parallelizable
- Initial use case: Machine Translation

Further use: *Kick-starts a new era of unsupervised representation learning.*

The Transformer



Source: Vaswani et al. (2017)

Advancing Word Embeddings

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

2013

Advancing Word Embeddings

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

February 2018 - ELMo

Guys from **AllenNLP** developed a bidirectionally contextual framework by proposing ELMo (**E**mbeddings from **L**anguage **Mo**dels; **Peters et al., 2018**).

Embeddings from this architecture are the (weighted) combination of the intermediate-layer representations produced by the biLSTM layers.

2013 | 01/2018 | 02/2018 | 06/2018

January 2018 - ULMFiT

The first transfer learning architecture (**U**niversal **L**anguage **M**odel **Fi**ne-**T**uning) was proposed by **Howard and Ruder (2018)**.
An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A **Unidirectional contextual** model since no biLSTMs are used.

June 2018 - OpenAI GPT

Radford et al., 2018 abandon the use of LSTMs. The combine multiple Transformer decoder block with a standard language modelling objective for pre-training.

Compared to ELMo it is just **unidirectionally contextual**, since it uses only the decoder side of the Transformer. On the other hand it is **end-to-end trainable** (cf. ULMFiT) and embeddings do not have to be extracted like in the case of ELMo.

Advancing Word Embeddings

2013 | 01/2018 | 02/2018 | 06/2018 | 10/2018

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

January 2018 - ULMFiT

The first transfer learning architecture (**U**niversal **L**anguage **M**odel **Fi**ne-**T**uning) was proposed by **Howard and Ruder (2018)**.
An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A **Unidirectional contextual** model since no biLSTMs are used.

February 2018 - ELMo

Guys from **AllenNLP** developed a bidirectionally contextual framework by proposing ELMo (**E**mbeddings from **L**anguage **Mo**dels; **Peters et al., 2018**).

Embeddings from this architecture are the (weighted) combination of the intermediate-layer representations produced by the biLSTM layers.

June 2018 - OpenAI GPT

Radford et al., 2018 abandon the use of LSTMs. The combine multiple Transformer decoder block with a standard language modelling objective for pre-training.

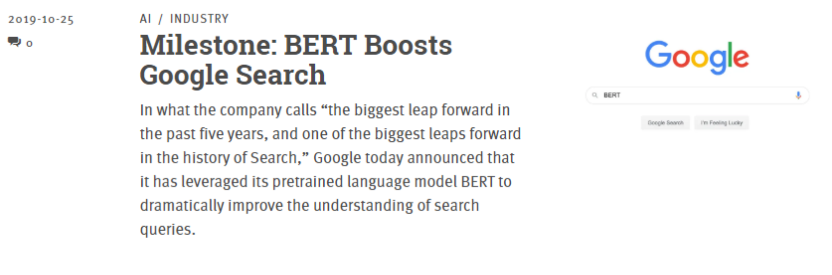
Compared to ELMo it is just **unidirectionally contextual**, since it uses only the decoder side of the Transformer. On the other hand it is **end-to-end trainable** (cf. ULMFiT) and embeddings do not have to be extracted like in the case of ELMo.

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

Since 2019:



2019-10-25

0

AI / INDUSTRY

Milestone: BERT Boosts Google Search

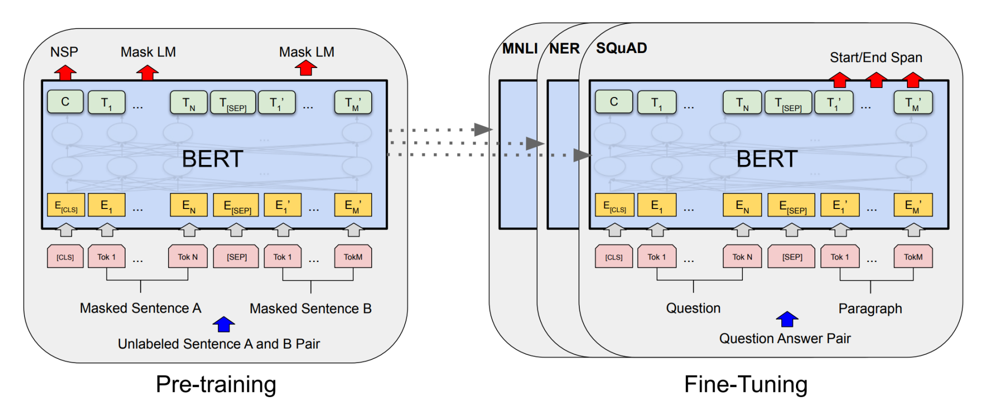
In what the company calls “the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search,” Google today announced that it has leveraged its pretrained language model BERT to dramatically improve the understanding of search queries.

Source: Synced

Corresponding blog post by Google:

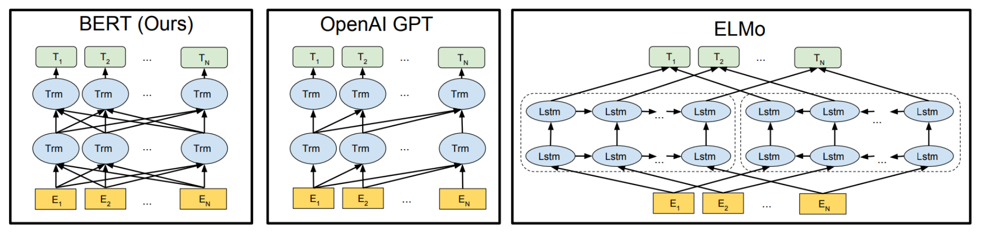
https://www.blog.google/products/search/search-language-understanding-bert/

Fine-tuning BERT



Source: Devlin et al. (2019)

GPT vs. ELMo vs. BERT



Source: Devlin et al. (2019)

Successors of BERT

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

10/2018

Successors of BERT

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

10/2018 → 02/2019

February 2019 - GPT2

Radford et al., 2019 massively scale up their GPT model from 2018 (up to 1.5 billion parameters). Despite the size, there are only smaller architectural changer, thus it remains a **unidirectionally contextual** model.

Controversial debate about this model, since OpenAI (at first) refuses to make their pre-trained architecture publicly available due to concerns about “malicious applications”.

Successors of BERT

10/2018 | 02/2019 | 06/2019

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

February 2019 - GPT2

Radford et al., 2019 massively scale up their GPT model from 2018 (up to 1.5 billion parameters). Despite the size, there are only smaller architectural changer, thus it remains a **unidirectionally contextual** model.

Controversial debate about this model, since OpenAI (at first) refuses to make their pre-trained architecture publicly available due to concerns about "malicious applications".

June 2019 - XLNet

Yang et al., 2019 design a new pre-training objective in order to overcome some weaknesses they spotted in the one used by BERT.

The use **Permutation Language Modelling** to avoid the discrepancy between pre-training and fine-tuning introduced by the artificial MASK token.

Successors of BERT

10/2018 | 02/2019 | 06/2019 | 07/2019

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

February 2019 - GPT2

Radford et al., 2019 massively scale up their GPT model from 2018 (up to 1.5 billion parameters). Despite the size, there are only smaller architectural changer, thus it remains a **unidirectionally contextual** model.

Controversial debate about this model, since OpenAI (at first) refuses to make their pre-trained architecture publicly available due to concerns about "malicious applications".

June 2019 - XLNet

Yang et al., 2019 design a new pre-training objective in order to overcome some weaknesses they spotted in the one used by BERT.

The use **Permutation Language Modelling** to avoid the discrepancy between pre-training and fine-tuning introduced by the artificial MASK token.

July 2019 - RoBERTa

Liu et al., 2019 concentrate on improving the original BERT architecture by (1) careful hyperaparameter tuning (2) abandoning the additional Next Sentence Prediction objective (3) increasing the pre-training corpus *massively*.

Other approaches now more and more concentrate on improving, down-scaling or understanding BERT. A new research direction called **BERTology** emerges.

Successors of BERT

10/2018 — 02/2019 — 06/2019 — 07/2019 — 10/2019

October 2018 - BERT

BERT (**Devlin et al., 2018**) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple **Transformer encoder** blocks.

BERT (and its successors) rely on the **Masked Language Modelling objective** during pre-training on huge unlabelled corpora of text.

February 2019 - GPT2

Radford et al., 2019 massively scale up their GPT model from 2018 (up to 1.5 billion parameters). Despite the size, there are only smaller architectural changer, thus it remains a **unidirectionally contextual** model.

Controversial debate about this model, since OpenAI (at first) refuses to make their pre-trained architecture publicly available due to concerns about "malicious applications".

June 2019 - XLNet

Yang et al., 2019 design a new pre-training objective in order to overcome some weaknesses they spotted in the one used by BERT.

The use **Permutation Language Modelling** to avoid the discrepancy between pre-training and fine-tuning introduced by the artificial MASK token.

July 2019 - RoBERTa

Liu et al., 2019 concentrate on improving the original BERT architecture by (1) careful hyperaparameter tuning (2) abandoning the additional Next Sentence Prediction objective (3) increasing the pre-training corpus *massively*.

Other approaches now more and more concentrate on improving, down-scaling or understanding BERT. A new research direction called **BERTology** emerges.

October 2019 - T5

T5 (**Raffel et al., 2019**) a complete **encoder-decoder** Transformer based architecture (**text-to-text transfer transformer**).

They approach transfer learning by transforming all inputs as well as all outputs to strings and fine-tuned their model simultaneously on data sets with multiple different tasks.

Now what does this all mean?

Key facts:

- *Super large* models applicable to a wide range of tasks
- Compute and data hungry, but:
  - pre-trained versions available
  - for a wide range of languages

Two exemplary use cases:

- Fake News Detection ▸ Guderlei & Aßenmacher (2020)
- Measuring customer centricity ▸ Lebmeier et al. (2021)

Task description: Stance detection of article body towards headline

Headline: Hundreds of Palestinians flee floods in Gaza as Israel opens dams	
Agree (AGR)	Hundreds of Palestinians were evacuated from their homes Sunday morning after Israeli authorities opened a number of dams near the border, flooding the Gaza Valley in the wake of a recent severe winter storm. [...]
Disagree (DSG)	Israel has rejected allegations by government officials in the Gaza strip that authorities were responsible for released storm waters flooding parts of the besieged area. "The claim is entirely false, and [...]" [...]
Discuss (DSC)	Palestinian officials say hundreds of Gazans were forced to evacuate after Israel opened the gates of several dams on the border with the Gaza Strip, and flooded at least 80 households. Israel has denied the claim as “entirely false”. [...]
Unrelated (UNR)	A Catholic priest from Massachusetts had been dead for 48 minutes before he was miraculously resuscitated. However, it is his description about God that is bound to spark a hot debate about the almighty. [...]

Fake News Detection ▸ Guderlei & Aßenmacher (2020)

Data:

- *Fake News Challenge Stage 1* (FNC-1): http://www.fakenewschallenge.org/
- Extension: *FNC-1 ARC*
  $\rightarrow$ Extends FNC-1 data by social media data

Goals:

- Compare performance of different pre-trained architectures
- Evaluate how much hyperparameter tuning is necessary
  - Learning rate
  - Learning rate schedule
  - Combinations of batch size & sequence length
- Experiment with different freezing techniques

Results:

	BERT		RoBERTa		DistilBERT		ALBERT		XLNet	
Metric	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC
$\mathbf{F_1}$-m	**70.18**	**72.20**	**78.18**	**78.19**	**72.11**	**73.59**	**59.80**	**65.01**	**75.00**	**75.57**
F_1-AGR	60.31	63.48	70.69	70.57	61.95	65.29	53.19	53.97	68.00	68.57
F_1-DSG	41.76	48.28	56.15	58.92	45.09	50.46	13.21	34.07	49.47	53.69
F_1-DSC	80.36	78.82	86.78	84.16	82.83	80.22	76.16	75.18	83.73	81.43
F_1-UNR	98.28	98.22	99.10	99.09	98.58	98.38	96.65	96.83	98.80	98.60

Table 4: Model performances with respect to class-wise F_1 as well as F_1-m in comparison for FNC-1 and FNC-1 ARC. For better readability we indicate the columns for FNC-1 ARC just with "+ ARC".

Main Takeaways:

- Important to not freeze to many layers
  - Freezing everything but the last layer yields *very* poor performance
  - Freezing none of the layers leads to longer fine-tuning times
  - Freezing the embedding layer yields good performance while saving time
- RoBERTa outperforms XLNet (at a lower computational expense)
  → Suspicion: *Sequence level task; XLNet cannot "play out" its strengths*
- Learning rate as most important hyperparameter
- Models relatively robust to changes in the other hyperparameters
- *Overall:* Generally strong performance with minimal hyperparameter tuning

Measuring customer centricity ▸ Lebmeier et al. (2021)

Problem statement:

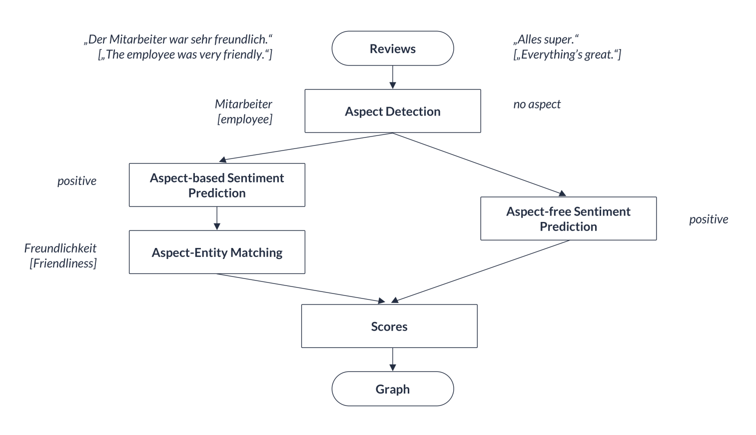
- Ubiquitous *direct* & *indirect* customer feedback,
  e.g. directly via e-mail or publicly available via comparison portals
- Oftentimes unstructured text (sometimes accompanied by star ratings)
- Vast amount of data prohibits a manual analysis of all feedback

→ *Goal:* Extract and visualize information automatically

The Project:

- Collaboration between Insaas and LMU via a student consulting project:
    - E. Lebmeier, N. Hou (M.Sc. students, Statistics)
    - K. Spann (Managing Director, Insaas)
    - C. Heumann, M. Aßenmacher (LMU)

Pipeline approach:

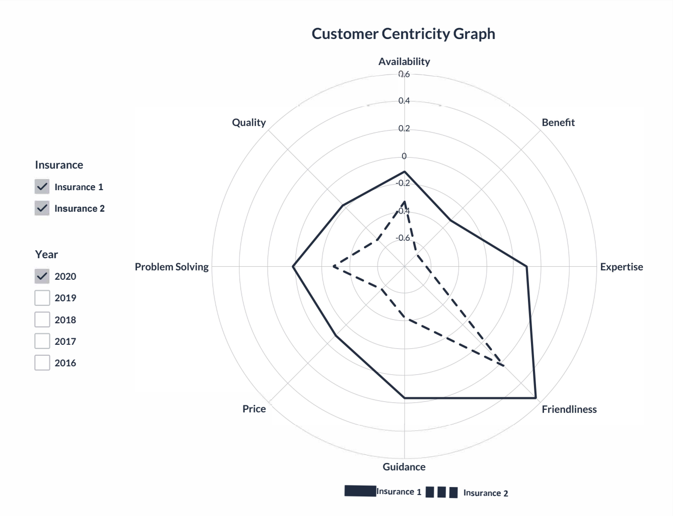


From raw reviews to the customer centricity graph

Measuring customer centricity ‣ Lebmeier et al. (2021)

The ingredients:

- Pre-trained German BERT models: ‣ German models @huggingface
- *Aspect detection:* ‣ German DistilBERT
- *Aspect-based Sentiment Prediction:* ‣ LCF-BERT
- *Aspect-Entity Matching:* Calculating the cosine similarity using FastText embeddings
- *Aspect-free Sentiment Prediction:* ‣ German DistilBERT



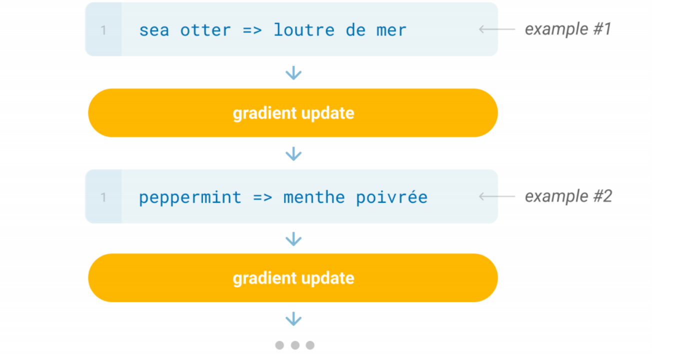
Exemplary customer centricity graph

Measuring customer centricity ▸ Lebmeier et al. (2021)

Demo version: ▸ Insaas vector

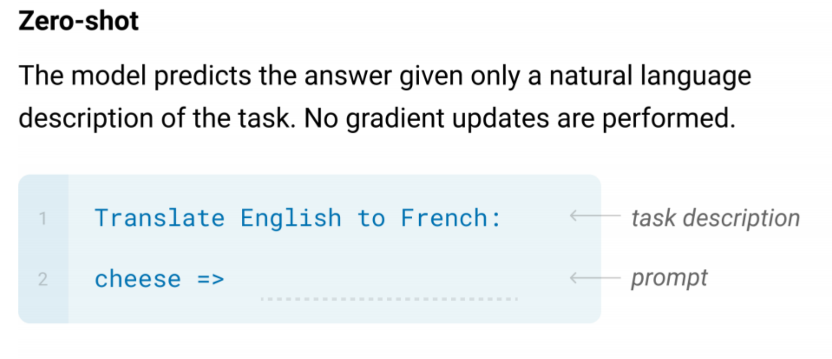
Outlook: GPT-3

"Classical" Pre-training–Fine-tuning paradigm:



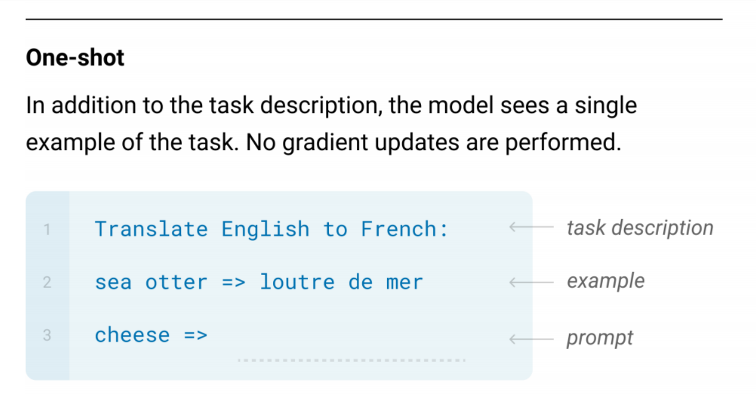
Source: Brown et al. (2020)

Zero-shot learning:



Source: Brown et al. (2020)

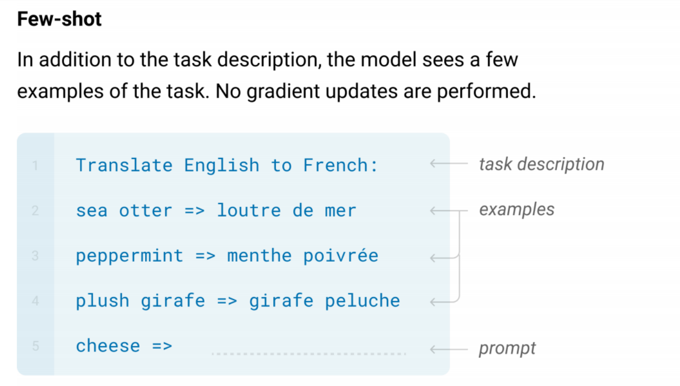
One-shot learning:



Source: Brown et al. (2020)

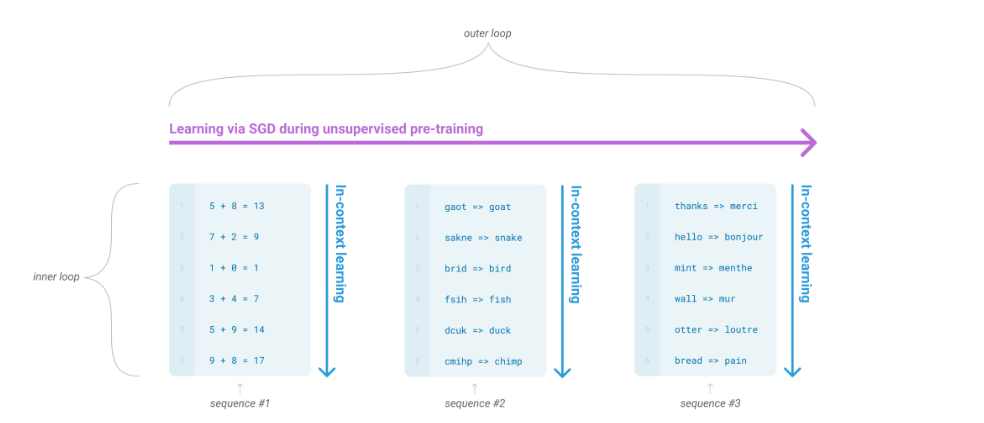
Outlook: GPT-3

Few-shot learning:



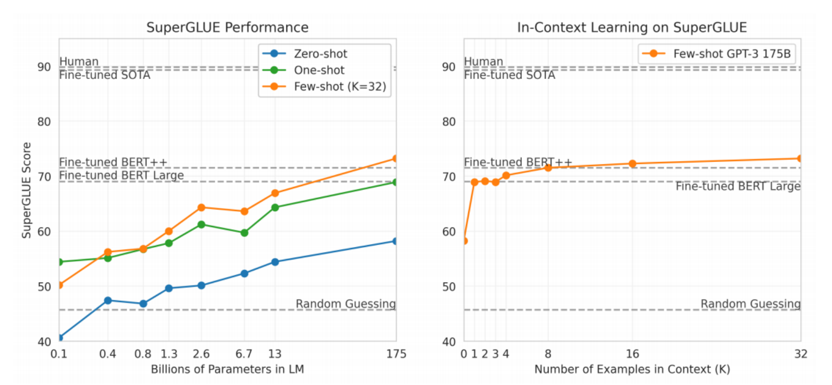
Source: Brown et al. (2020)

Outlook: GPT-3



Source: Brown et al. (2020)

Performance:



Source: Brown et al. (2020)