Neural Embeddings

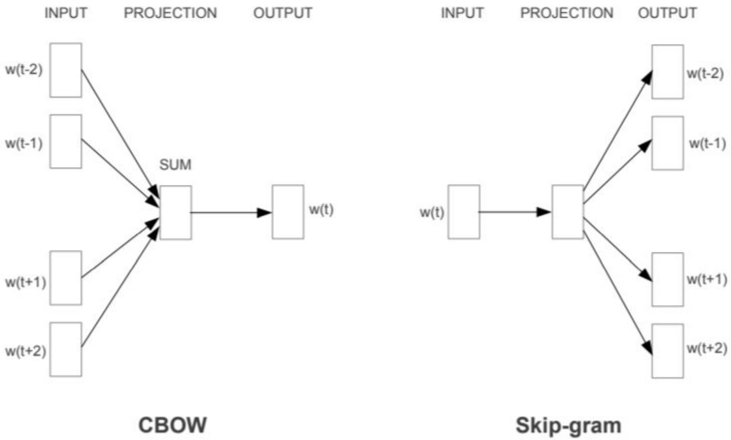
Trends in Language Representation since Word2Vec

Philipp Hager
Research Assistant, Department of Marketing & Management, SDU
Data Scientist, Blinks Labs GmbH
phil@sam.sdu.dk

April 2021

Word2Vec - 2013

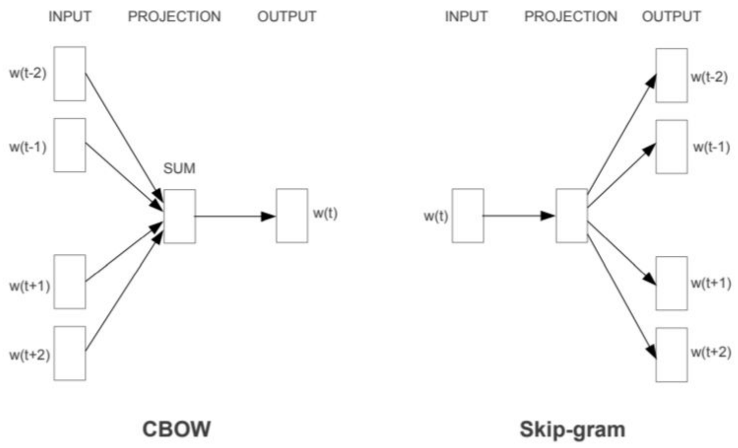
- Published in 2013 by Mikolov et al. at Google [1, 2, 3]
- Vectors can encapsulate
    - **Syntactic relationships** (kind, kindly, kindest)
    - **Semantic relationships** (brother, sister, family)
    - **Arithmetics** (Queen - Woman ≈ King)



Word2Vec - 2013

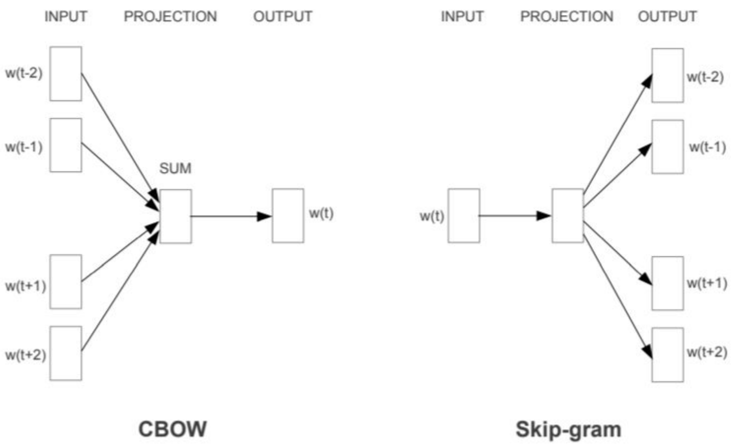
- Published in 2013 by Mikolov et al. at Google [1, 2, 3]
- Vectors can encapsulate
    - **Syntactic relationships** (kind, kindly, kindest)
    - **Semantic relationships** (brother, sister, family)
    - **Arithmetics** (Queen - Woman ≈ King)

Later studies show that learned relations can pick up **biases** such as gender stereotypes. [4], [5]



Word2Vec - 2013

- Published in 2013 by Mikolov et al. at Google [1, 2, 3]
- Vectors can encapsulate
    - **Syntactic relationships** (kind, kindly, kindest)
    - **Semantic relationships** (brother, sister, family)
    - **Arithmetics** (Queen - Woman ≈ King)



Later studies show that learned relations can pick up **biases** such as gender stereotypes. [4], [5]

Properly tuned matrix factorization methods like **SVD** and **LSA** can achieve similar performance to Word2Vec. Data > Model [6]

FastText - 2016

Problems of Word2Vec (and GloVe)

- Each word has its own embedding, morphology of is not used: *improve, improvement*
- No embeddings for unknown words
  - Typos: *improvment*
  - Slang: *heeey*
  - Compound nouns: *cutting-edge*

FastText

- Published in 2016 by Bojanowski et al. at Facebook [7, 8]
- Uses **character-level embeddings** to represent a word

FastText - 2016

Idea

- Represents a word as the **sum of its character n-grams** (length 3-6)
- N-grams at the start and the end of a word are treated differently by adding pre-/suffix
- Grammatical variations of a word will share most of their n-grams

	3-grams	**4-grams**	**5-grams**	**6-grams**
	^lo	^lov	^love	^love$
	lov	love	love$	
^love$	ove	ove$		
	ev$			

FastText - 2016

query	tiling	tech-rich	english-born	micromanaging	eateries	dendritic
sisg	tile flooring	tech-dominated tech-heavy	british-born polish-born	micromanage micromanaged	restaurants eaterie	dendrite dendrites
sg	bookcases built-ins	technology-heavy .ixic	most-capped ex-scotland	defang internalise	restaurants delis	epithelial p53

Nearest Neighbors to the unknown word “english-born” in FastText and Word2Vec [7]

From Words to Sentences

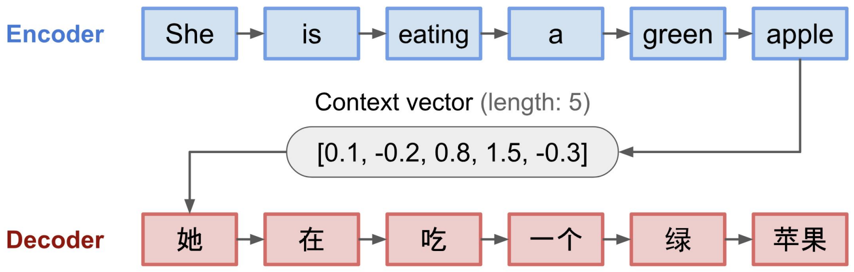
Sentence Representations

- **(Weighted) Averaged Word Embeddings:**
  - Simple but surprisingly strong baseline [9]
- **Extensions of Word Embedding Models:**
  - Doc2Vec: Word2Vec extension [10]
  - Sent2Vec: FastText extension [11]
- **RNNs and Transformers:**
  - ELMo [12]
  - Universal Sentence Encoder [13]
- **Non-neural**: Topic modelling, Bag of Words...

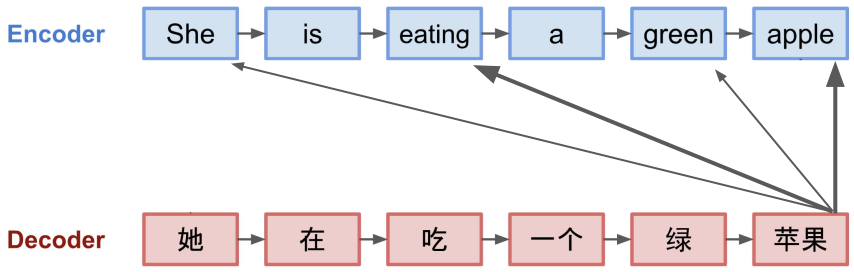
Attention - 2015

Attention

- Published by Bahdanau et al. for Machine Translation [14]
- Neural network layer that allows the network to **attend specific parts of the NN input**



Machine Translation using a traditional RNN Encoder-Decoder architecture
[Source]

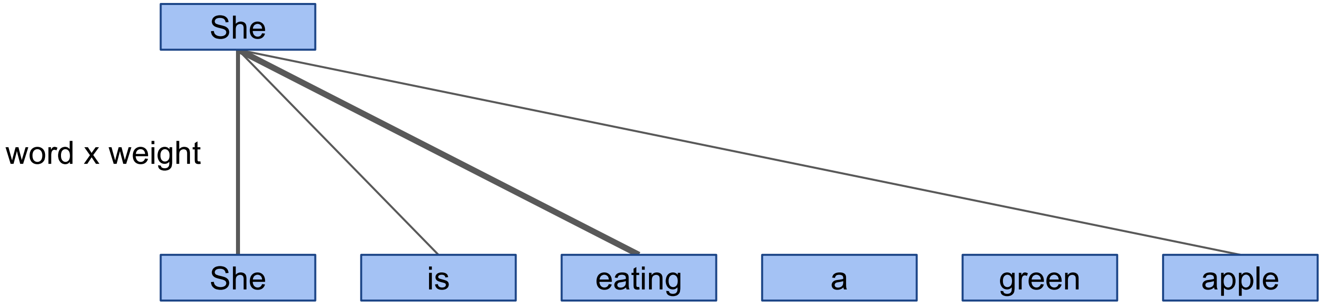


Attention allows the Decoder network to access outputs from the Encoder
[Source]

Self-Attention - 2016

Self-Attention

- First published by Cheng et al. in 2016 [15]
- The network can only attend to part of the input sequence itself
- Useful for many use-cases including: Summarization, classification, translation
- *“Weighted averaging of input vectors and the weights are learned by the network”*

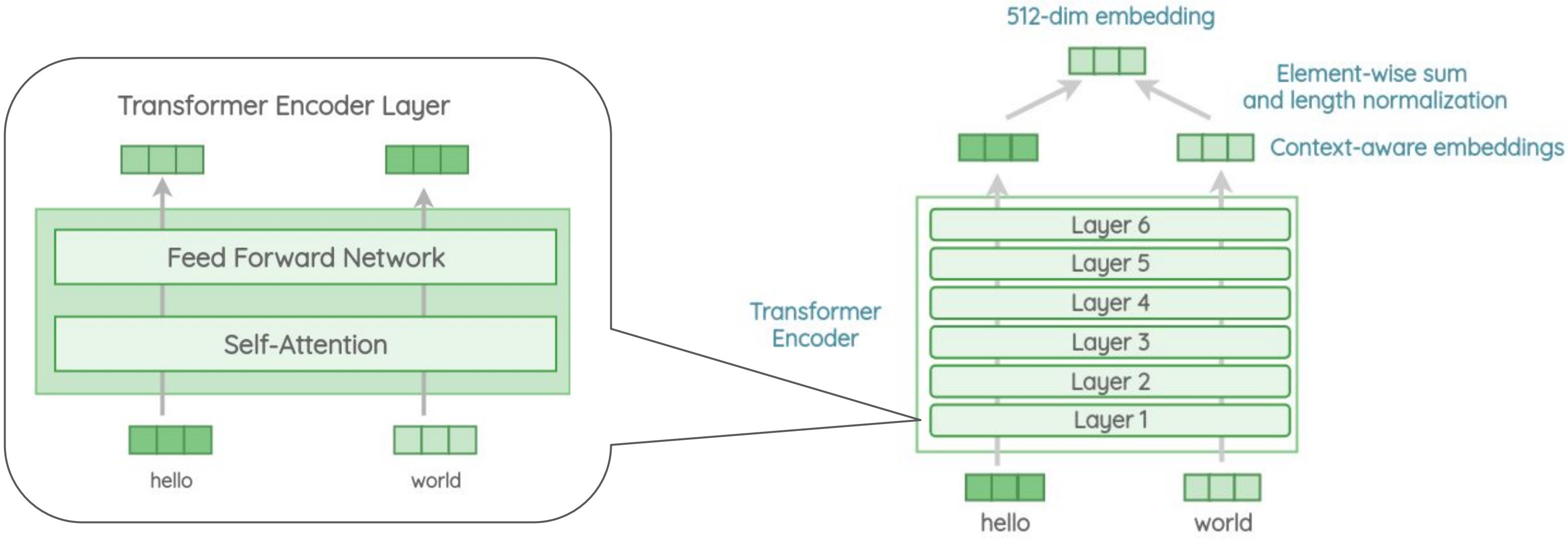


Transformers - 2017

Transformer

- Published by Vaswani et al. at Google [16]
- Encoder / Decoder model for Machine Translation
- Stacks layers of self-attention and feed-forward layers
- **Context-aware:** Embeddings change based on the surrounding words
- **Positional encoding:** Word order matters (like in RNNs)
- **Explainability:** Attention can be inspected and visualized

Transformers - 2017

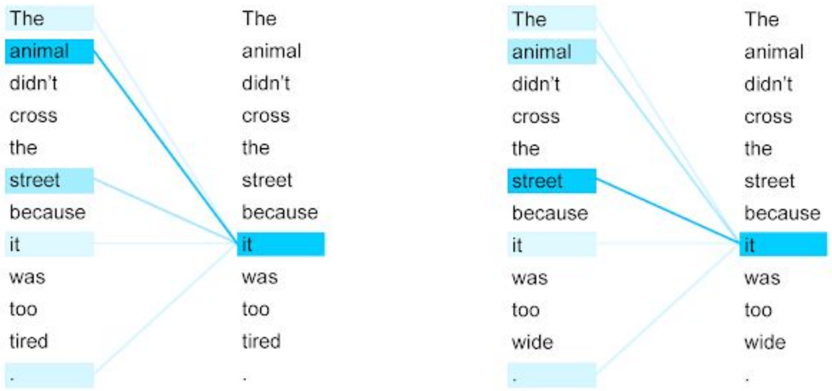


[Source]

Transformers - 2017

[Source]

Transformers - 2017



Self-Attention visualizing **coreference resolution** of the word ***"it"*** inside a Transformer. [Source]

Large Pre-trained Language Models

Fine-Tuning

- Train models with billions of parameters
- Train NN models on vast amounts of data
- Tune model on small-task specific datasets

Criticism, Bender et al. [17]

- Environmental impact
- Financial barrier
- Larger datasets -> larger risks?

HuggingFace Library 🤗

https://huggingface.co/models

Year	Model	# of Parameters	Dataset Size
2019	BERT	340,000,000	16GB
2019	DistilBERT	66,000,000	16GB
2019	ALBERT	223,000,000	16GB
2019	XLNet	340,000,000	126GB
2020	ERNIE-Gen	340,000,000	16GB
2019	RoBERTa	355,000,000	161GB
2019	MegatronLM	8,300,000,000	174GB
2020	T5-11B	11,000,000,000	745GB
2020	T-NLG	17,000,000,000	174GB
2020	GPT-3	175,000,000,000	570GB
2020	GShard	600,000,000,000	–
2021	Switch-C	1,570,000,000,000	745GB

Size of State-Of-The-Art Transformer models over the last years. [17]

Multilingual Word Embeddings

Representing words or documents from **multiple languages** in the **same embedding space**

Applications [18]

- Machine translation
- Cross-lingual information retrieval
- Transfer trained models across languages

Design Differences [18]

- **Alignment:** Word, sentence, document level
- **Comparability:** Exact translations or roughly comparable data

Multilingual Word Embeddings

Dataset Examples

(a) Word, par. (b) Word, comp. (c) Sentence, par. (d) Sentence, comp. (e) Doc., comp.

Different types of alignment-level and levels of comparability [18]

Comp. = Comparable data
Par. = Exact, parallel translations

NLP in Production

A Content-based Recommender System Case Study

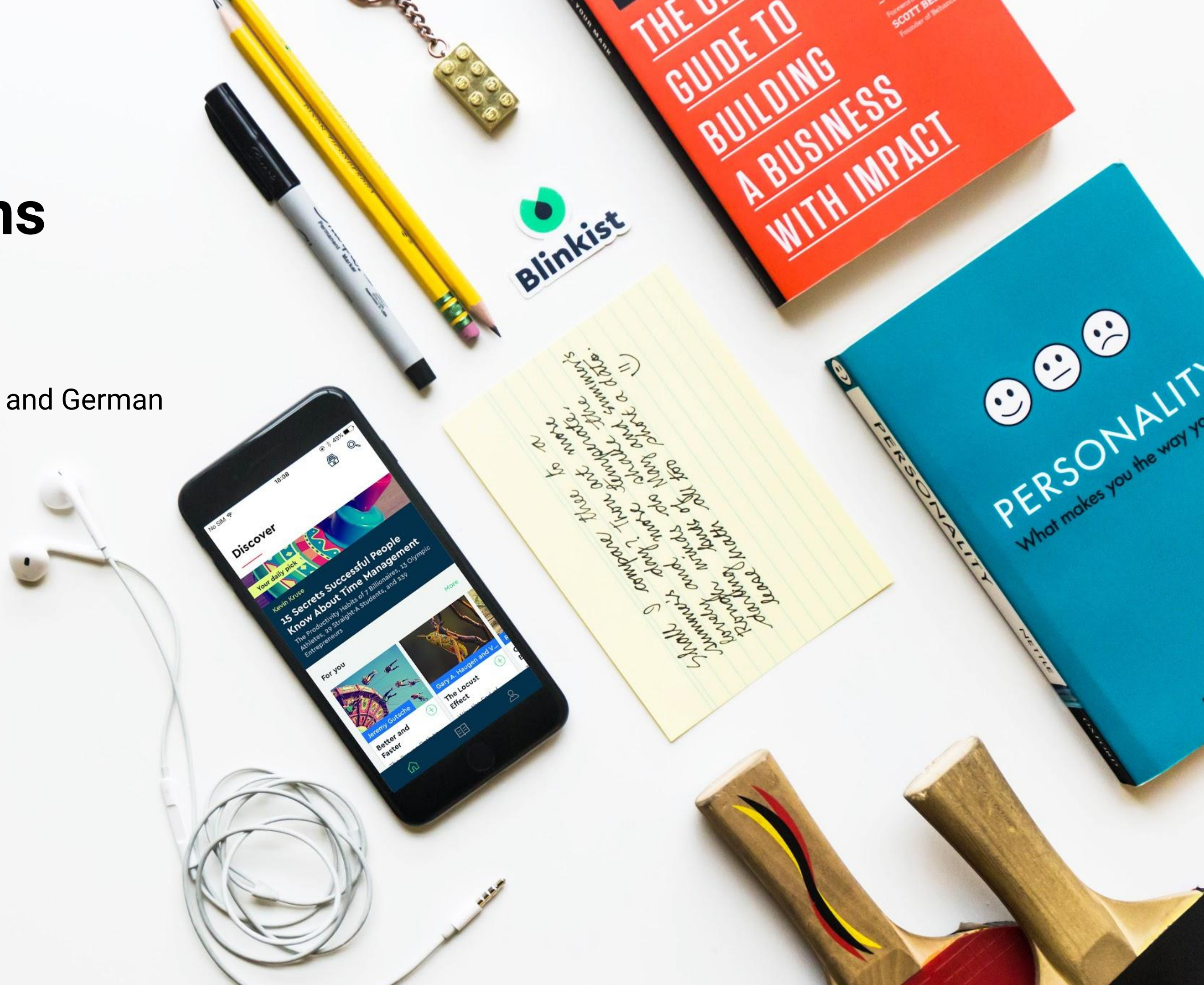
Philipp Hager
Research Assistant, Department of Marketing & Management, SDU
Data Scientist, Blinks Labs GmbH

April 2021

Using NLP for Recommendations

Use Case

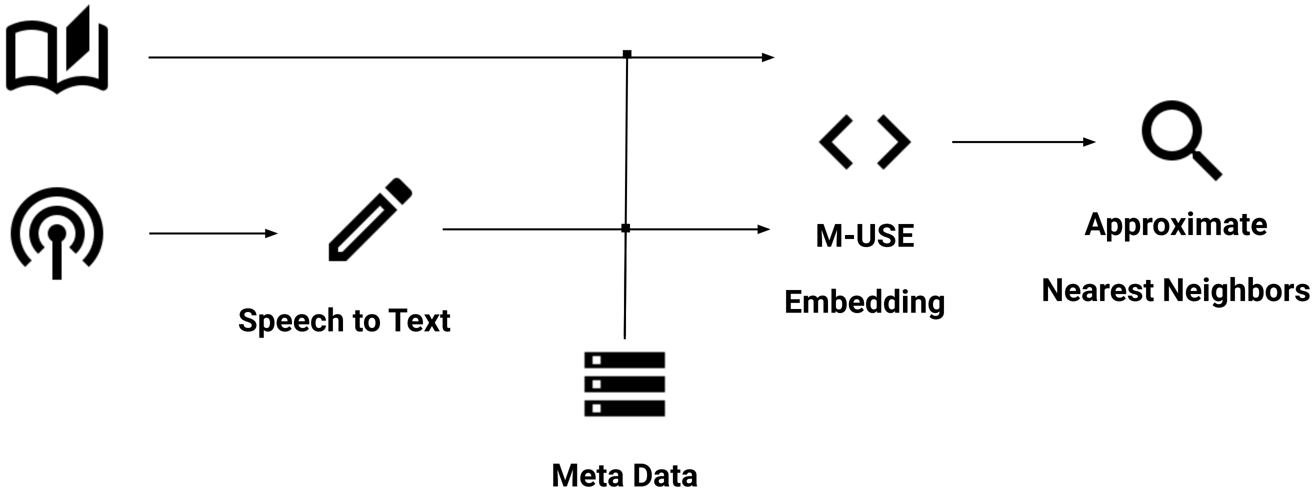
- **Multilingual content:** English and German
- **Multiple content formats:**
  - Books
  - Podcasts
  - Audiobooks
  - ...



Content-Based Recommendation

Use Case

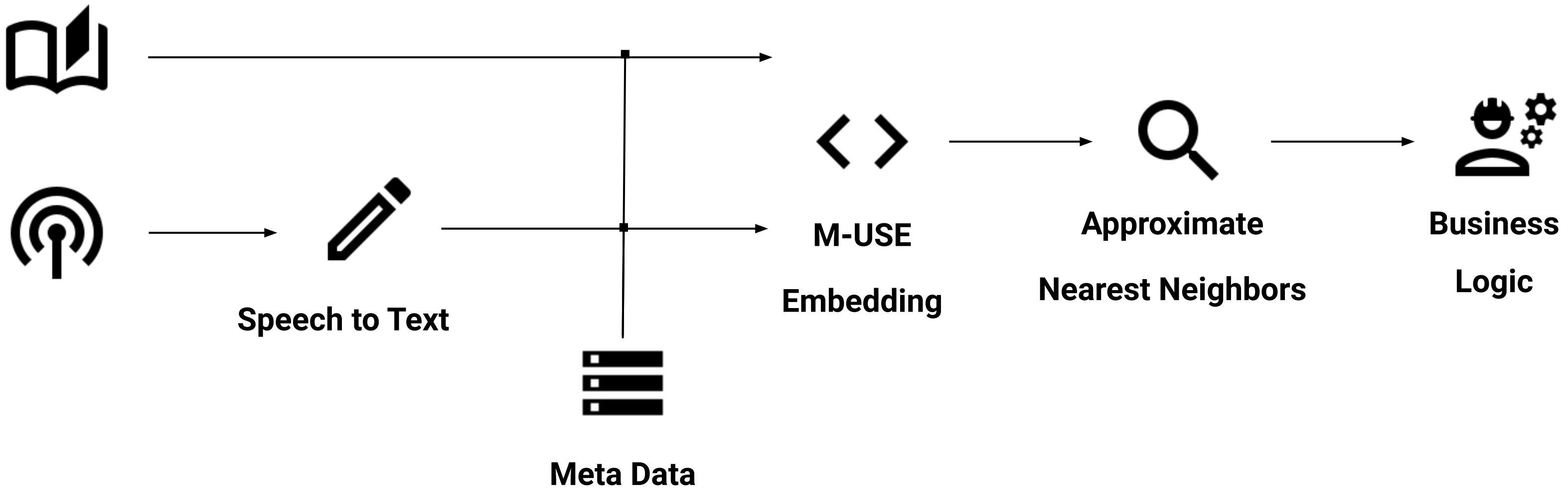
- **Multilingual content:** In English and German
- **Multiple content formats:** Books, Podcasts, Audiobooks, etc.



Content-Based Recommendation

Use Case

- **Multilingual content:** In English and German
- **Multiple content formats:** Books, Podcasts, Audiobooks, etc.

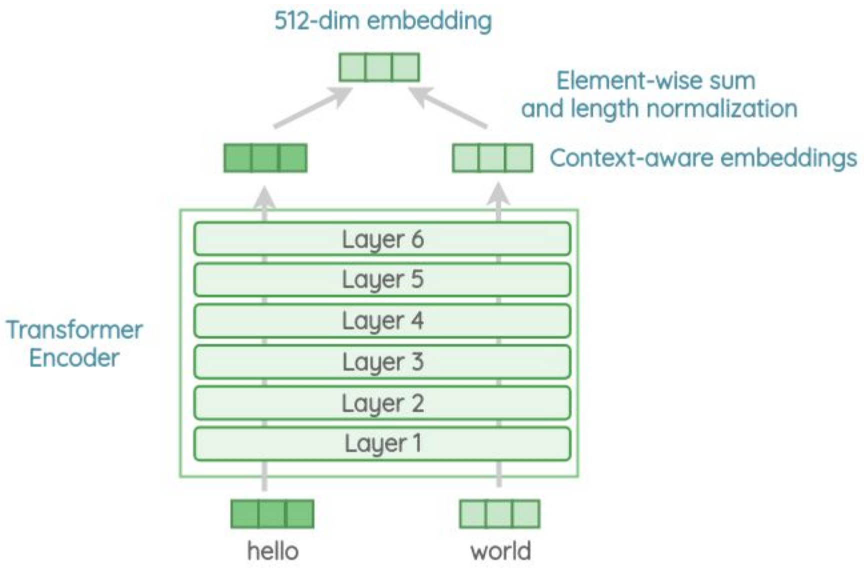


Multilingual Universal Sentence Encoder - 2019

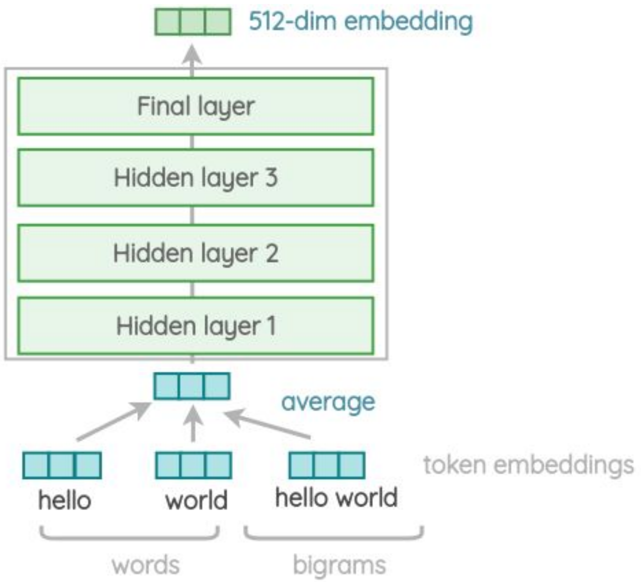
M-USE

- Proposed by Google in 2019 [13, 19, 20]
- **Contextual, multilingual, character-level, sentence embeddings**
- Trained on **16 languages**
- **Architectures:**
  - **Transformer-based:** Contextual Sentence Embeddings (higher accuracy)
  - **Deep Averaging Network**: Simple Feed-Forward Network (higher speed)
  - **(CNN)**
- **Multi-Task Learning:** Train the same embedding to solve multiple downstream tasks

Universal Sentence Encoder - 2018

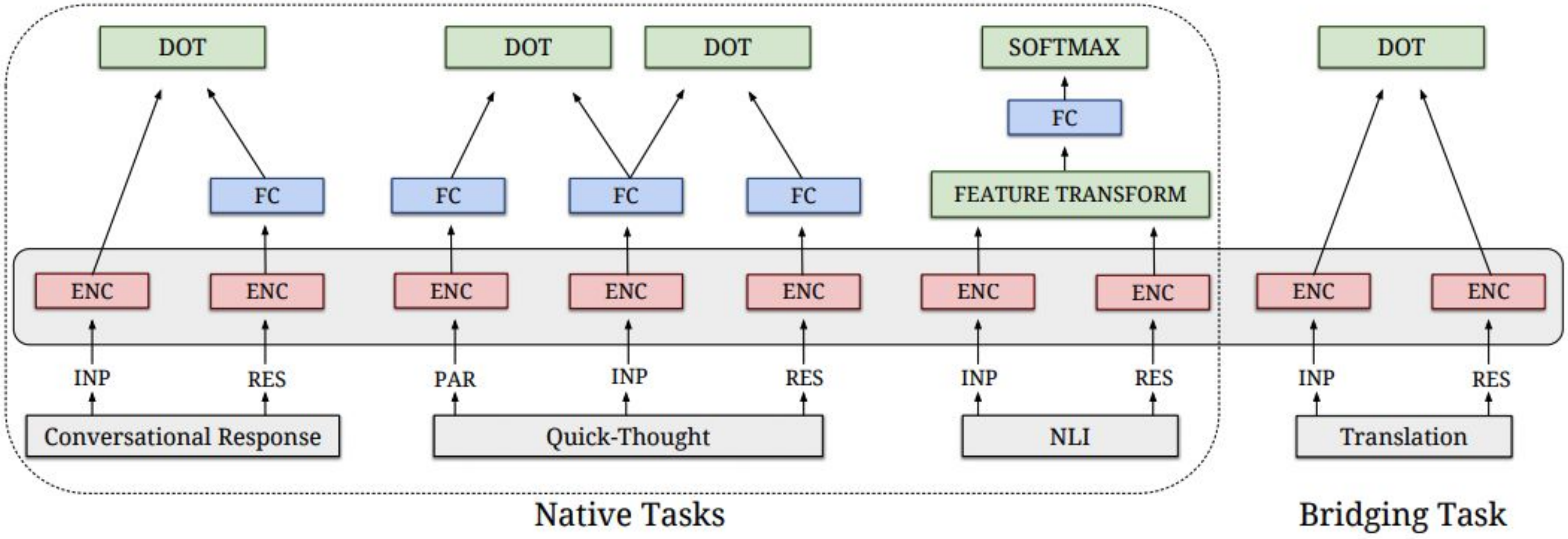


Encoder part of the Transformer [Source]



Deep Averaging Network (DAN) [Source]

Multilingual Universal Sentence Encoder - 2019



Multi-Task learning setup of M-USE [20]

Demo

Lessons from Data Science at a Startup

Clean Data > Approach

Speed > State-Of-The-Art

Balance MVP work and maintainable code

Prefer simple and robust approaches

References

[1] Tomás Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Poster at ICLR 2013.

[2] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed representations of words and phrases and their compositionality. NIPS 2013.

[3] Tomás Mikolov, Wen-tau Yih, Geoffrey Zweig. Linguistic regularities in continuous space word representations. NAACL-HLT 2013.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS 2016.

[5] Hila Gonen, Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. NAACL 2019.

[6] Omer Levy, Yoav Goldberg, Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the ACL, 2015, volume 3, pages 211–225.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomás Mikolov. Enriching word vectors with subword information. TACL 2017.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomás Mikolov. Bag of Tricks for Efficient Text Classification. EACL 2017.

[9] Sanjeev Arora, Yingyu Liang, Tengyu Ma. A Simple but thought-to-beat Baseline for Sentence Embeddings. ICLR 2017.

References

[10] Quoc Le and Tomás Mikolov. Distributed representations of sentences and documents. ICML 2014.

[11] Matteo Pagliardini, Prakhar Gupta, Martin Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. NAACL 2018.

[12] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. Deep contextualized word representations. NAACL 2018.

[13] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder for English. EMNLP 2018.

[14] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015.

[15] Jianpeng Cheng, Li Dong, Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. EMNLP 2016.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is All you Need. NIPS 2017.

[17] Emily Bender, Timnit Gebru, Angelina McMillan, Shmargaret Majorand Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜, ACM FAccT 2021.

[18] Sebastian Ruder, Ivan Vulić, Anders Søgaard. A Survey of Cross-lingual Word Embedding Models. Journal of Artificial Intelligence Research, 2019, volume 65, pages 569-631.

References

[19] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego , Steve Yuan, Chris Tar, Yun-hsuan Sung, Ray Kurzweil. Multilingual Universal Sentence Encoder for Semantic Retrieval. ACL 2019.

[20] Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. Repl4NLP@ACL 2019.