School of CIT –
Social Computing
Research Group

Technical
University
of Munich

Advanced Natural Language Processing
CIT4230002

Prof. Dr. Georg Groh
Carolin Schuster, M.Sc.

Lecture 2.1
Embeddings:
Analysis & Applications

Lecture Outline

- Quick Recap (NLP1) on Semantics and Word2Vec
- Overview on Embeddings: Types, Levels, Training Tasks
- Methods for analyzing Embeddings
- Applications => Focus on Topic Modeling

Embeddings:
Recap NLP1 & Overview

Recap NLP1 | Lexical Semantics

The study of word meaning

- Word relatedness / association

semantic field: set of semantically related items
e.g. mammals: rodents, bats, primates, etc.

- Distinction between word senses

WordNet Examples for “bat”

- S: (n) **bat**, chiropteran (nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate)
- S: (n) **bat** (a club used for hitting a ball in various games)

- Relations between word senses

Synonymy vs. antonymy, hyponymy vs. Hypernymy

- Affective meaning of words (connotation): sentiment, opinions, evaluations

Recap NLP1 | Distributional Semantics

"you shall know a word by the company it keeps" – Firth (1957)

$\Rightarrow$ words that occur in similar contexts tend to have similar meanings

Distributional semantics:

- meaning of a word is computed from the distribution of words around it
- word embedding: a vector representation of a word
- semantic similarity of words => similarity of word vectors
- Simplest version: co-occurrence matrices of words as vectors

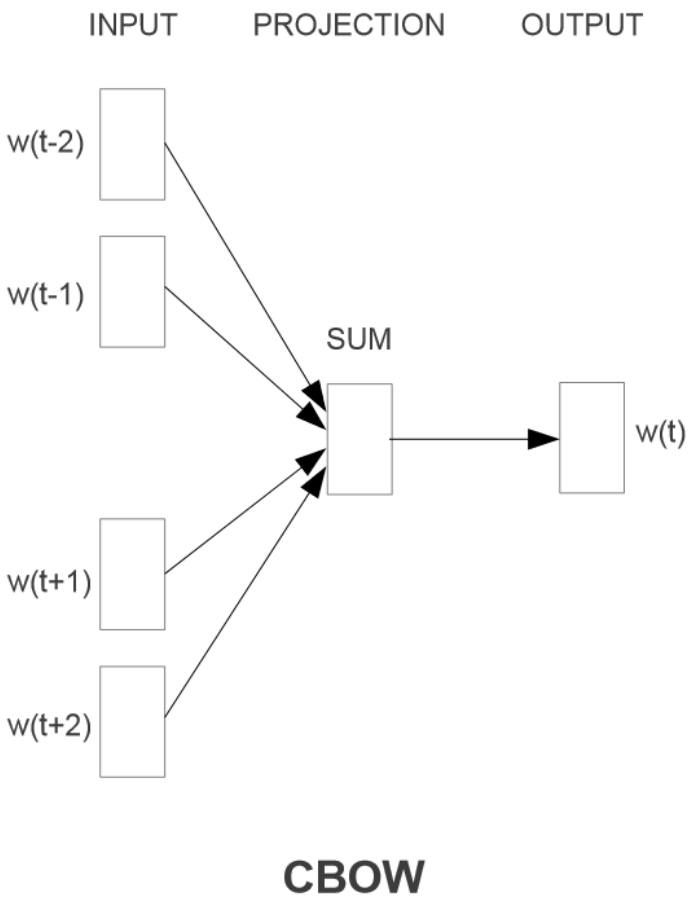
	the	bat	is	flying	bird	bites	sings
bat	2	2	1	1	0	1	0
bird	2	0	1	1	2	0	1

Recap NLP1 | Word2Vec

Word2Vec

- Training a shallow neural network to **predict the current word from surrounding words** (CBOW)
  or predict context words from the current word (Skip-gram)

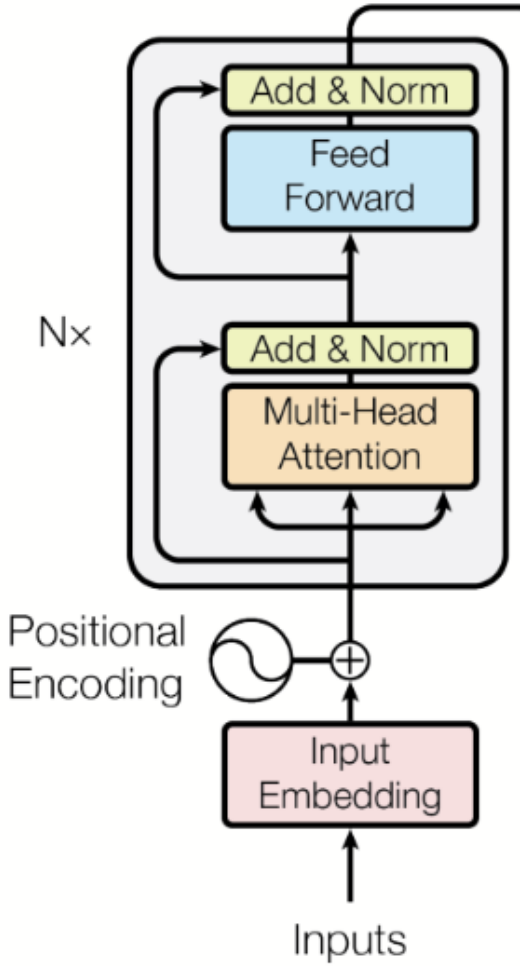
- Word embedding:
  Neural network weights



Recap NLP1 | Contextual Embeddings

Contextual Embeddings from LLMs

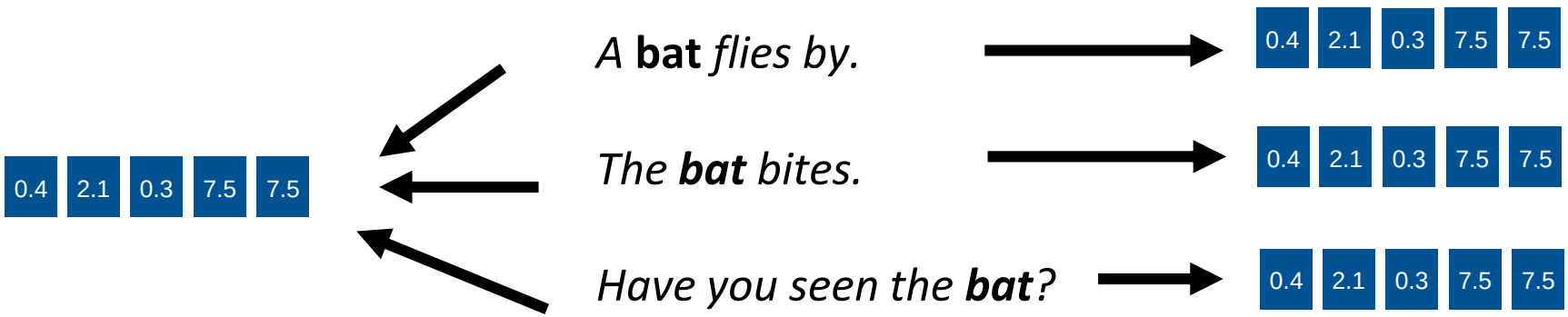
- Deeper neural networks produce contextualized representations
  => meaning of a word in its context
- Often a transformer encoder, e.g. BERT
- Word Embedding:
  Neural network **weights for a word in a specific context**



Static vs. Contextual Embeddings

Static Embeddings:
Word2Vec, GloVe, FastText, etc.

Contextual Embeddings:
BERT, ElMo, DeBERTa etc.



Static vs. Contextual Embeddings

Static Embeddings

Simple representation

- One vector per word
- Context not considered
- No representation of polysemous words

Fast training, does not require as much data

$\Rightarrow$ Can be trained from scratch

Contextual Embeddings

More expressive representation of language:

- Unlimited possible representations for each word
- Represent words in their contexts
- Distinguish word senses

Training requires huge amounts of data and computing resources

$\Rightarrow$ Finetuning for adaption

BUT Pre-trained models introduce their own biases

=> Even more challenges regarding interpretability

Current paradigm of Subword Tokenization

- Byte-Pair Encoding (BPE): iteratively merging frequent pairs of tokens until desired vocabulary size is reached
- Algorithms such as BPE depend on a training dataset
- Resulting tokenization may not be optimal for every application and domain, important words might be split-up!

Pooling Token Embeddings

Strategies for higher-order embeddings:

- Embedding of special tokens, e.g. [CLS] for BERT (sentence representations)
- Averaging sub-embeddings => best simple strategy for obtaining a full word representation from subwords
- Finetuning with extended domain-specific vocabulary:

  New word embeddings can be initialized by similar words or averaging of subtoken embeddings

Another dimension of pooling concerns the layers, e.g. we might average the embedding over the last four layers (e.g. common for semantic shift analysis).

Training Tasks for Contextual Embeddings

Token Level Training
characters/subwords/words

Training Representations for Sequences of Tokens
phrases, sentences, documents

Language Modeling (only one-directional context)

Masked Language Modeling

Replaced Token Detection

Whole Word Masking

Span Boundary Objective (SpanBERT)

Cosine Similarity

Contrastive Learning

Triplet Loss

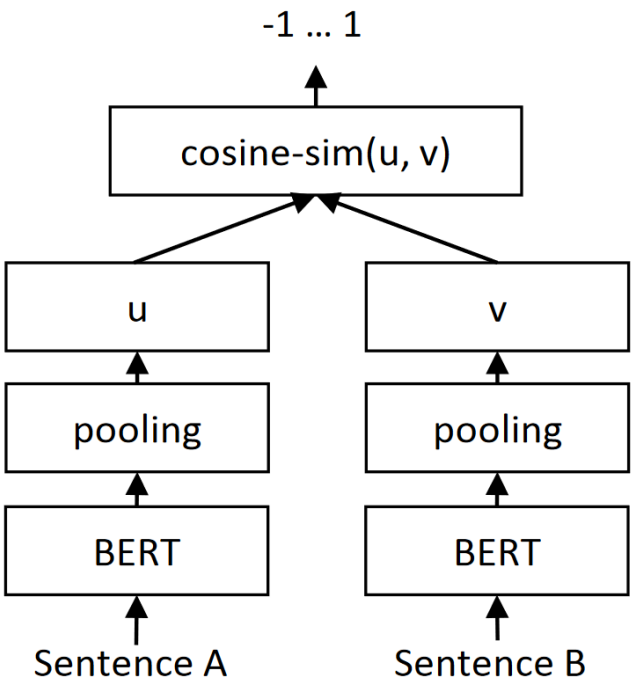
SimCSE

PromptBERT

Sequential Denoising Auto-Encoder (TSDAE)

Example: Sentence-BERT Training

Cosine Similarity Loss



Triplet Loss
(Contrastive Learning)

- Euclidian distance
- minimize:

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$$

s_a anchor sentence embedding
s_p positive sentence embedding
s_n negative sentence embedding

Specialized and Extended Contextual Embeddings

- Domain-specific: BERTweet, SciBERT, PatentBERT, BioBERT
- Monolingual: CamemBERT, FinBERT, TiBERT, IndoBERT,…
- Multilingual: mBERT, XLM-RoBERTa, multiligual SBERT, multilingual Universal Sentence Encoder
- Multi-Modal: SpeechBERT, CLIP, BLIB
- Knowledge Enhanced: e.g. KnowBERT with WordNet and Wikipedia

Adapting Contextual Embeddings

Adaption with Training

- Continued Pre-Training on in-domain data
- Finetuning with task-specific data, e.g. STS, NLI

Instruction Tuning (INSTRUCTOR)

- Embeddings based on both text input and the task description

  e.g. “Represent the review comment for classifying the emotion as positive or negative”
- Adaption without any training

Embedding Benchmark

MTEB: Massive Text Embedding Benchmark

- 8 Tasks, 56 datasets, 117 languages (Bi-Text Mining)
- Sentence and paragraph level tasks
- Limitations: Task and language imbalance, no word level tasks

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)
1	voyage-large-2-instruct			1024	16000	68.28	81.49	53.35	89.24
2	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54
3	gte-Qwen1.5-7B-instruct					67.34	79.6	55.83	87.38
4	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42	86.87
5	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61	87.16

...

⋮

- Obtaining a general idea of embedding capacity, but there is no one size fits all solution

Analysis of Contextual Embeddings

Why analyze contextual embeddings?

What did the neural networks learn?

What information is encoded?

- About language
- About the world

Is there any harmful information encoded?

What can the embeddings be used for?

How to analyze contextual embeddings?

Dimensionality Reduction

Clustering

„Probing“ with Classifiers

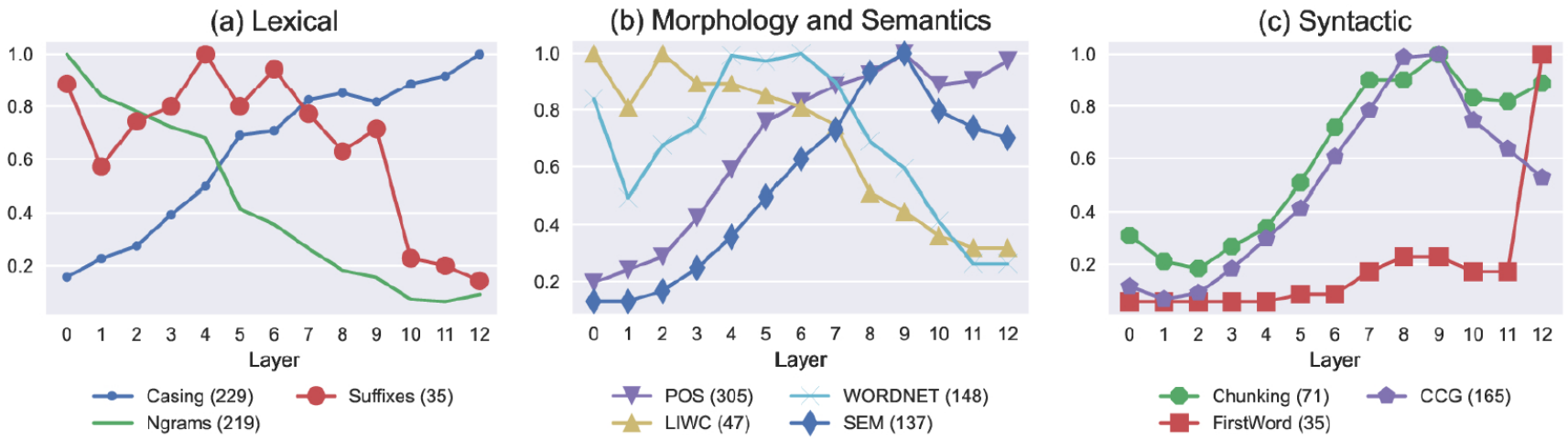
Inspiration from the Social Sciences:

- Association Tests
- Semantic Differentials

Dimensionality Reduction

PCA on BERT sentence embeddings reveal a moral direction (Schramowski, 2022)

Clustering

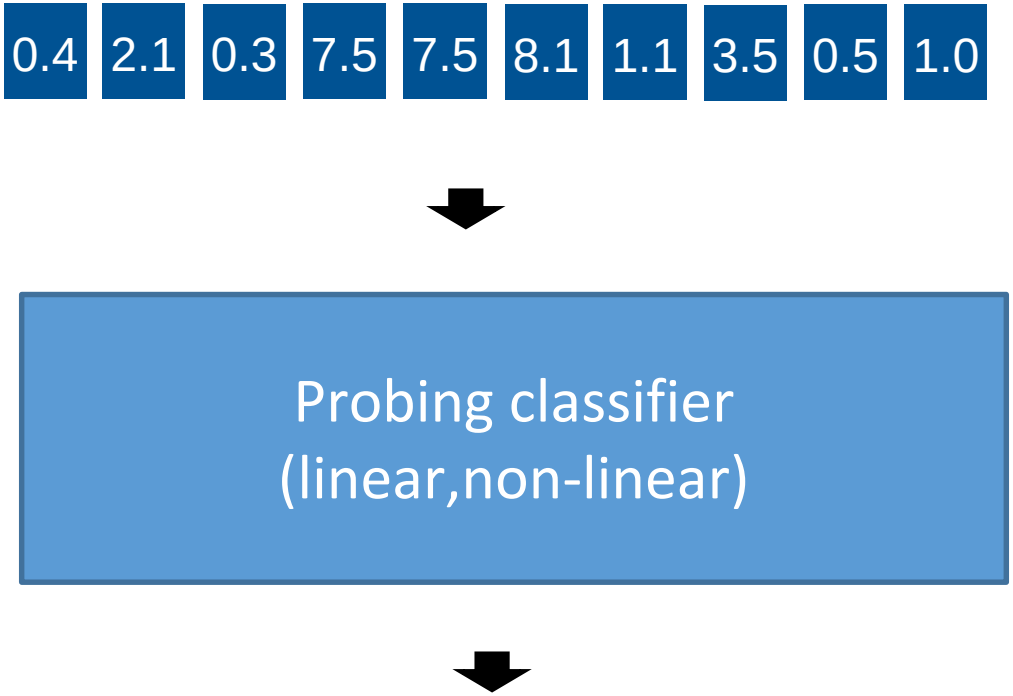


Layer-wise alignment of pre-defined concepts in BERT (Dalvi, 2022)

(a) (b) (c)

Example Concepts

Probing | Predicting a property of interest



"Estimating the mutual information between representations and a property of interest"

Property of interest:
(Part of speech, semantic roles/relations, etc.)

Probing | Dis(Advantages)

Advantage: Flexibility

- Any type of embedding
- Any property

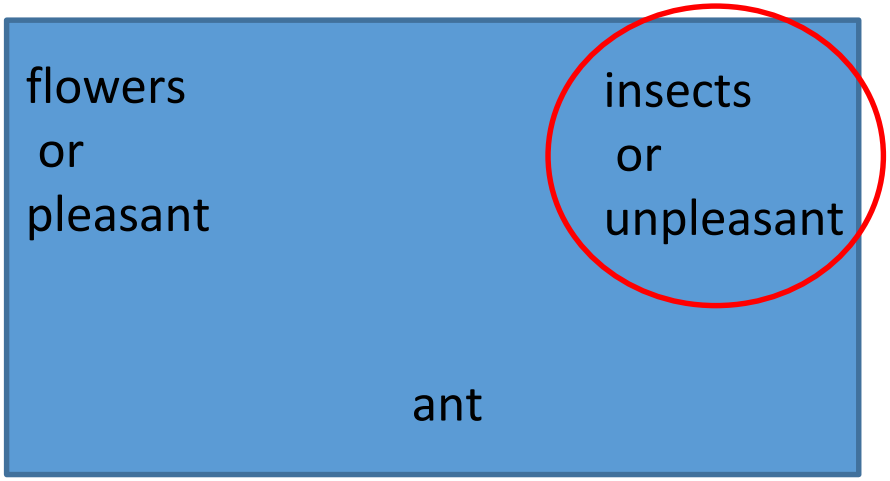
Caveats

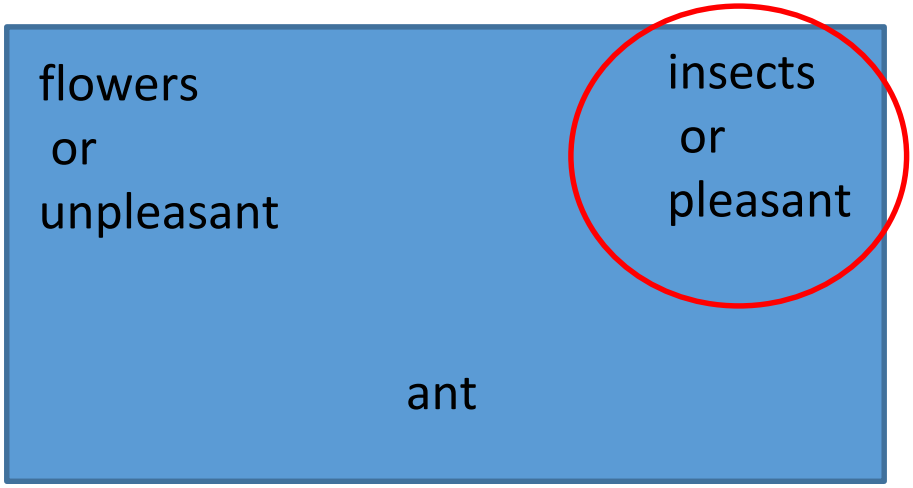
- Accuracy-complexity trade-off: What classifier to use?
- Correlations with property of interest

=> Designing accurate probing task is difficult

Association Tests | Inspiration

Implicit Association Test (Greenwald, 1998)

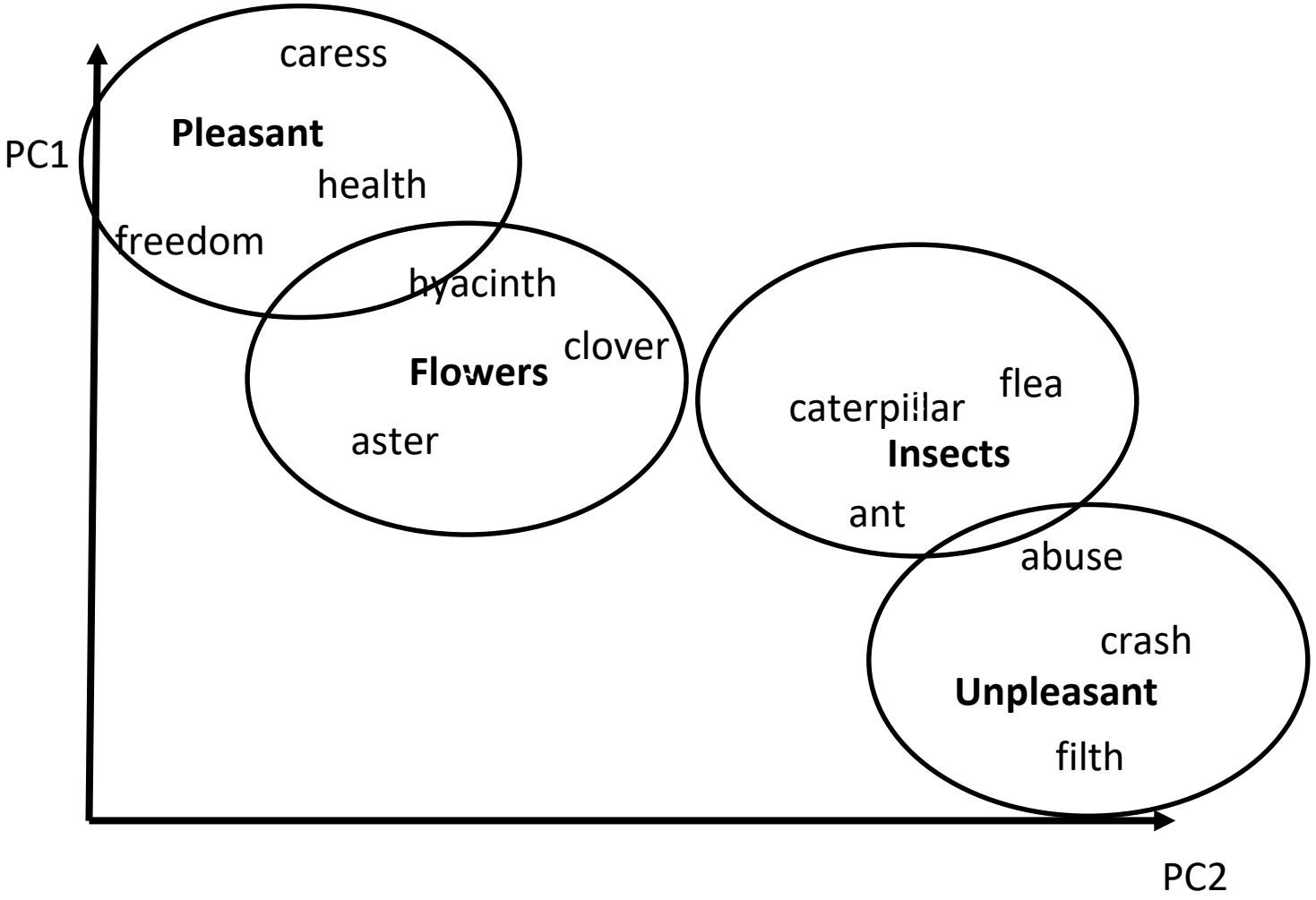




- Measuring **differential association** of two target concepts with an attribute in a choice task
- Response time as implicit measure of association
- Faster response time when concepts are associated (e.g. flowers and pleasant)

Association Tests | Transfer to Embedding Space

Measuring relative distances of two target concepts to an attribute



Association Tests | WEAT

Word Embedding Association Test (Caliskan, 2017)

- Cosine similarity to measure associations

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

$$s(w,A,B) = \text{mean}_{a \in A} \cos(\vec{w},\vec{a}) - \text{mean}_{b \in B} \cos(\vec{w},\vec{b})$$

  - *X target set 1 (Insects)*
  - *Y target set 2 (Flowers)*
  - *A attribute set 1 (Pleasant Words)*
  - *B attribute set 2 (Unpleasant Words)*

Findings => ***Word embeddings contain human-like biases***

- Morally neutral bias (insects vs. flowers)
- **Racial and gender bias**
- Example: Math and Arts
  - *X target set 1 (math, algebra, etc.)*
  - *Y target set 2 (poetry, art, etc.)*
  - *A attribute set 1 (male, man, boy, etc.)*
  - *B attribute set 2 (female, woman, girl, etc.)*

Association Tests | CEAT

Contextualized Embedding Association Test (Guo, 2021)

- Sampling contexts from a reference corpus

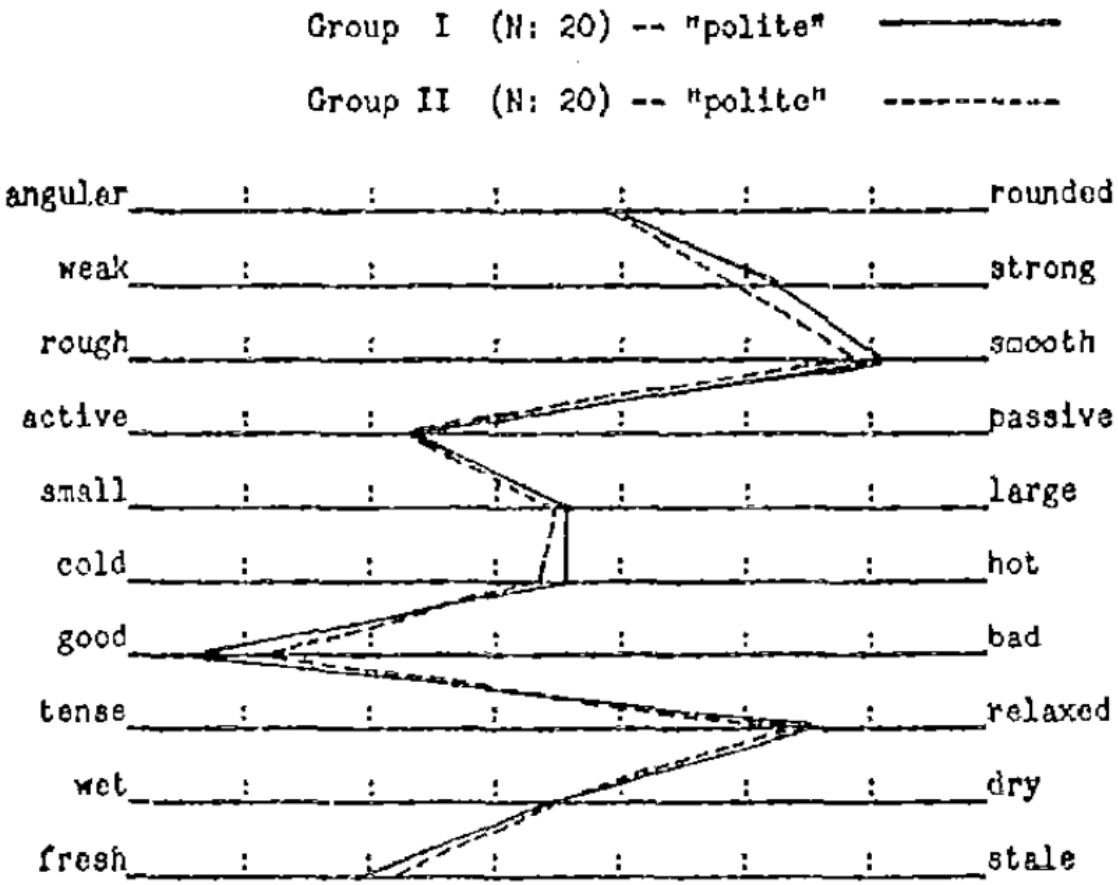
Test			ELMo		BERT		GPT		GPT-2	
			d	p	d	p	d	p	d	p
C1:	Flowers/Insects	random	1.40	$< 10^{-30}$	0.97	$< 10^{-30}$	1.04	$< 10^{-30}$	0.14	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	1.35	$< 10^{-30}$	0.64	$< 10^{-30}$	1.01	$< 10^{-30}$	0.21	$< 10^{-30}$
C2:	Instruments/Weapons	random	1.56	$< 10^{-30}$	0.94	$< 10^{-30}$	1.12	$< 10^{-30}$	-0.27	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	1.59	$< 10^{-30}$	0.54	$< 10^{-30}$	1.09	$< 10^{-30}$	-0.21	$< 10^{-30}$
C3:	EA/AA names	random	0.49	$< 10^{-30}$	0.44	$< 10^{-30}$	-0.11	$< 10^{-30}$	-0.19	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	0.47	$< 10^{-30}$	0.31	$< 10^{-30}$	-0.10	$< 10^{-30}$	0.09	$< 10^{-30}$
C7:	Math/Arts	random	0.64	$< 10^{-30}$	0.41	$< 10^{-30}$	0.24	$< 10^{-30}$	-0.01	$< 10^{-2}$
	Male/Female terms	fixed	0.71	$< 10^{-30}$	0.20	$< 10^{-30}$	0.23	$< 10^{-30}$	-0.14	$< 10^{-30}$
C8:	Science/Arts	random	0.33	$< 10^{-30}$	-0.07	$< 10^{-30}$	0.26	$< 10^{-30}$	-0.16	$< 10^{-30}$
	Male/Female terms	fixed	0.51	$< 10^{-30}$	0.17	$< 10^{-30}$	0.35	$< 10^{-30}$	-0.05	$< 10^{-30}$
C9:	Mental/Physical disease	random	1.00	$< 10^{-30}$	0.53	$< 10^{-30}$	0.08	$< 10^{-29}$	0.10	$< 10^{-30}$
	Temporary/Permanent	fixed	1.01	$< 10^{-30}$	0.40	$< 10^{-30}$	-0.23	$< 10^{-30}$	-0.21	$< 10^{-30}$

Table excerpt from Guo (2021). Two rows per test: Completely random samples vs. identical sentences across all neural language models.

Polar Embeddings | Inspiration

Semantic Differentials (Osgood, 1952)

- Measuring the meaning of concepts with polar scales



Polar Embeddings | POLAR Framework

POLAR framework (Mathew, 2020)

- Embedding polar opposites (antonyms) to identify an **interpretable subspace**

Step 1: Definition of the polar embedding space with a set of polar opposites

$$\mathbb{P} = \{(p_z^1, p_{-z}^1), (p_z^2, p_{-z}^2), \ldots, (p_z^N, p_{-z}^N)\}$$

$$\overrightarrow{dir_1} = \overrightarrow{\mathbb{W}^a_{p_z^1}} - \overrightarrow{\mathbb{W}^a_{p_{-z}^1}}$$

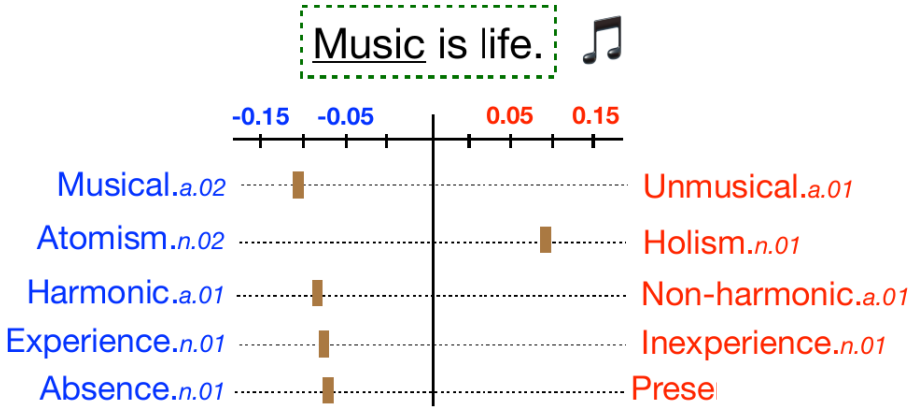
Step 2: Projection of a word to the interpretable embedding space

$$dir^T \overrightarrow{\mathbb{E}_v} = \overrightarrow{\mathbb{W}^a_v}$$

$$\overrightarrow{\mathbb{E}_v} = (dir^T)^{-1} \overrightarrow{\mathbb{W}^a_v}$$

Extension to Contextual Embeddings: SensePOLAR (Engler, 2022)

- Sense embeddings instead of word embeddings
- **Context examples for word senses from dictionaries** (e.g. WordNet)
- Context examples for projection



	SQuAD 1.1		SQuAD 2.0	
Metric	Base	SensePOLAR	Base	SensePOLAR
EM	86.92	86.85↓ 0.07%	80.88	81.06↑ 0.22%
F1	93.15	93.12↓ 0.03%	83.87	83.89↑ 0.02%

Table 2: Results of fine-tuned BERT embeddings and with SensePOLAR transformed embeddings on the

- Polar embeddings are interpretable with similar performance

Polar embeddings | Extensions and Limitations

Extension to broader concepts

- Word list for each pole
- Example: Stereotype dimensions (Fraser, 2021)
  Word lists for warmth
  high: *friendly, warm, pleasant, etc.*
  low: *cold, repellent, disliked, etc.*

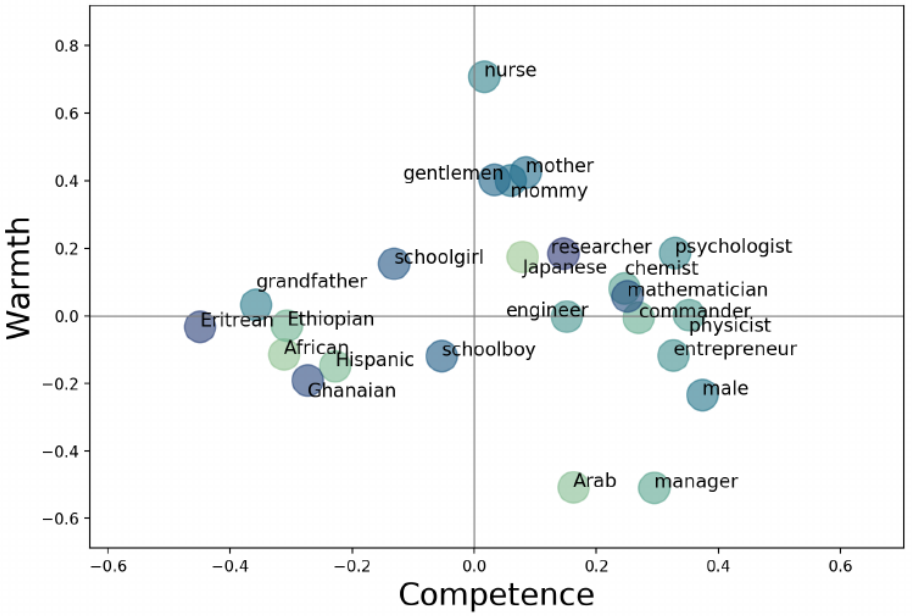


Figure 1: Validating known stereotypes.

Limitations

- Requirement of Opposites
- Dependency on dictionaries (antonym relations, word lists)
- Requirement of quality context examples (potential source of bias)

Analysis of Contextual Embeddings | Opportunities

Opportunities

- Getting a sense of what LLMs learn (+ Reality check of what LLMs are)
- Bias Measurement, BUT measurements in embeddings and downstream tasks may not give the same results
- De-Biasing, same caveat as above
- Optimizing embedding spaces
- Using embeddings to study (digital) society

Analysis of Contextual Embeddings | Caveats

Caveats

Distributive properties of embeddings make comprehensive analysis difficult

- High dimensionality + distribution across layers
- Anisotropy, artifacts (rogue dimensions)
- Entanglement of meaning, facts, syntax..
- High variability depending on context (what context examples should be employed in the analysis?)

Applications of Embeddings
Topic Modeling

Applications of Embeddings

Pre-Trained Contexual Embeddings beyond LLM Analysis

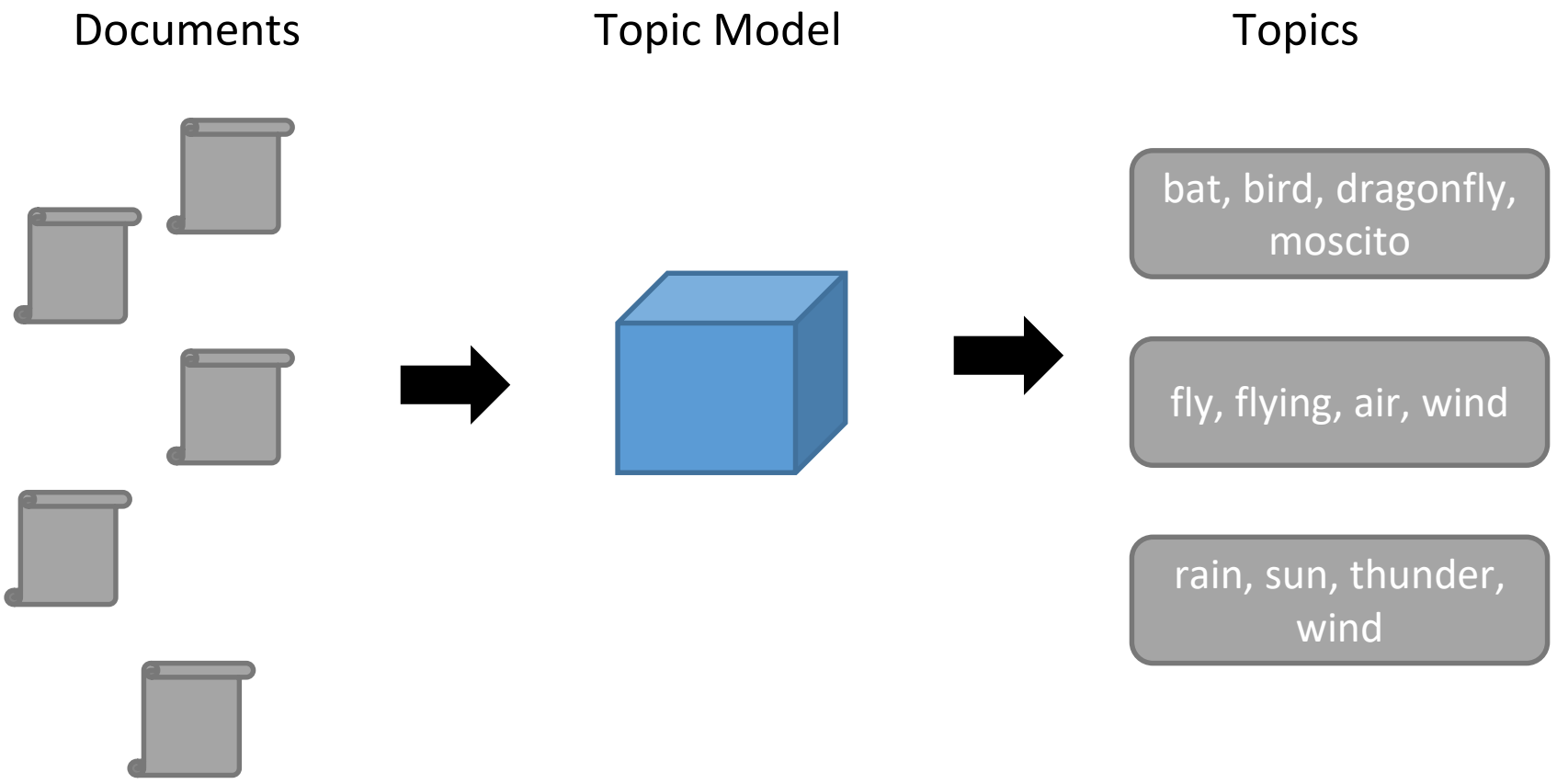
- Semantic search
- RAG (will be covered later in the lecture)
- Exploratory data analysis
- Studying meaning and semantic shift of words
- Document clustering (sentence/document embeddings)
- **Topic modeling** <= Focus of this lecture and corresponding tutorial

Still static embeddings are popular with researchers and practioners

- Fast to train from scratch (no bias from pre-training data)
- Studying bias of a dataset NOT bias of a LLM

Topic Modeling | Introduction

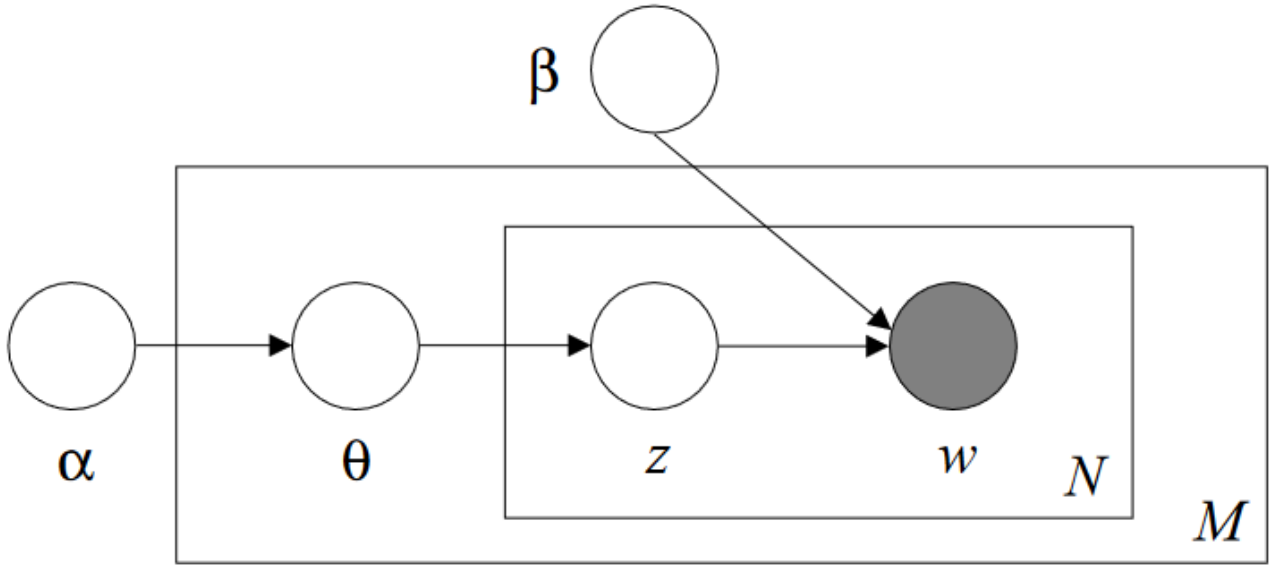
Topic Modeling
= Identifying abstract themes in a collection of documents



Topic Modeling | LDA

Before neural topic models there was **Latent Dirichlet Allocation** (LDA) (Blei, 2003)

- Generative probabilistic model
- Documents are represented as random mixtures over latent topics
- Topics are characterized by a **distribution over words**



M denotes the number of documents

N is number of words in a given document (document *i* has N_i words)

α is the parameter of the Dirichlet prior on the per-document topic distributions

β is the parameter of the Dirichlet prior on the per-topic word distribution

θ_i is the topic distribution for document *i*

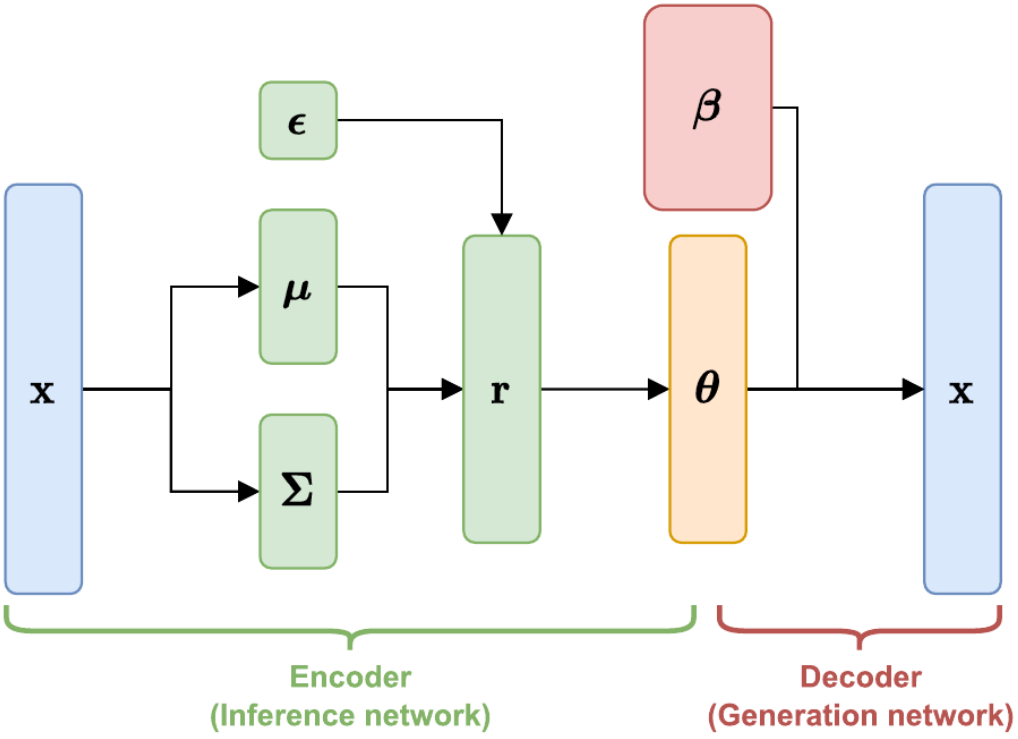
φ_k is the word distribution for topic *k*

z_{ij} is the topic for the *j*-th word in document *i*

w_{ij} is the specific word.

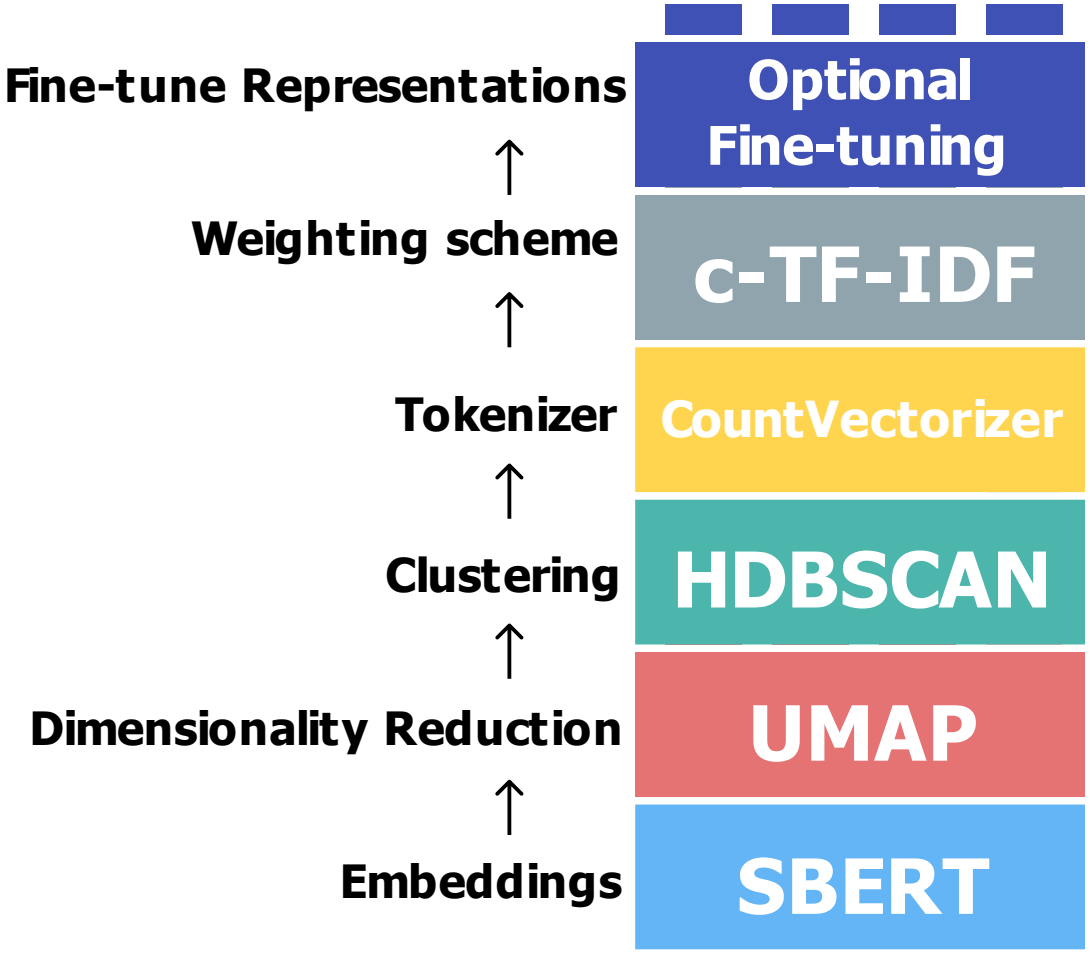
Neural Topic Modeling with Variational Auto-Encoders (e.g. Miao, 2017)

- Compression of the document to topic distribution Θ by an encoder
- Reconstruction of the document by a decoder network

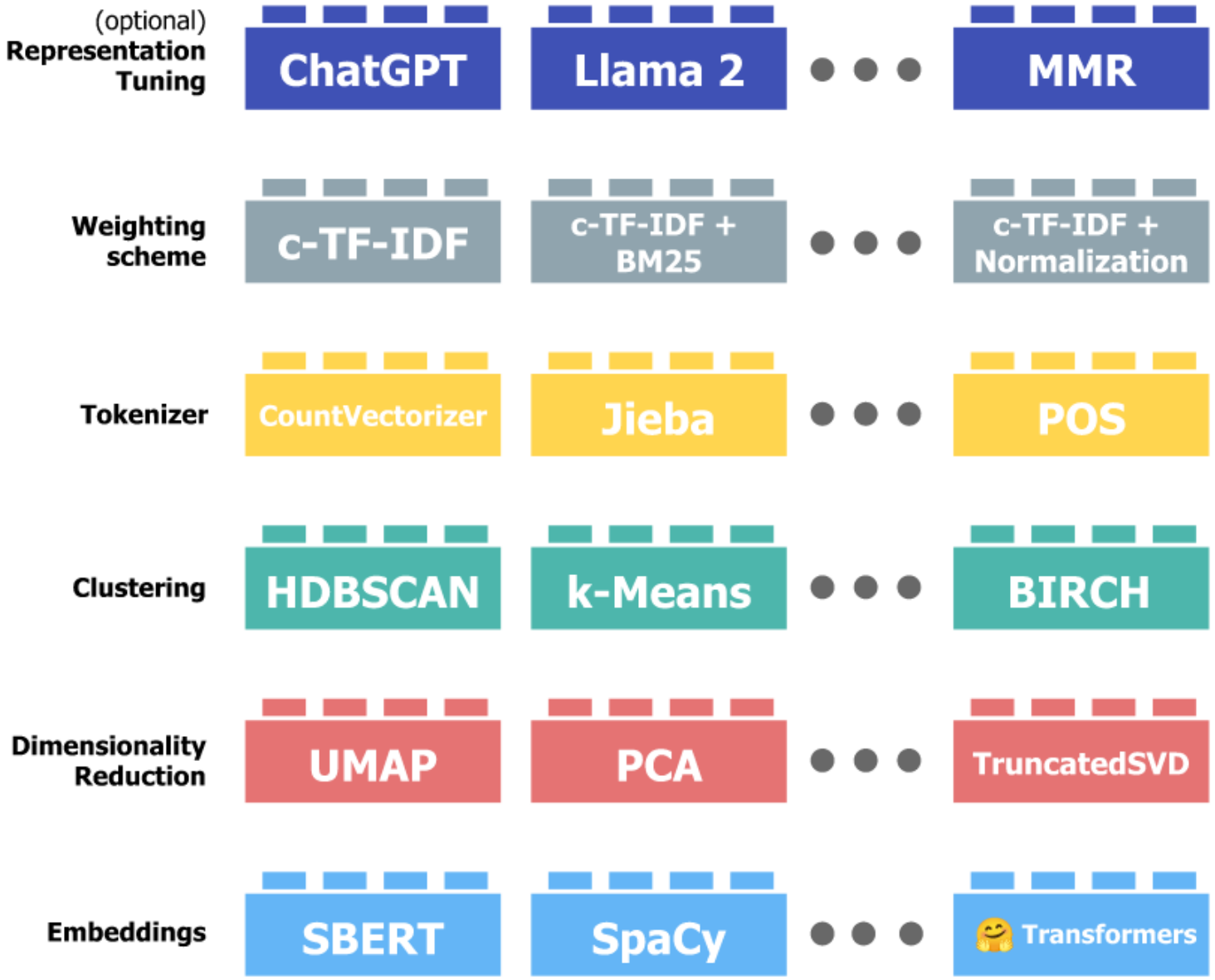


https://link.springer.com/content/pdf/10.1007/s10462-023-10661-7.pdf

BERTopic (Grootendorst, 2022): Modular approach to topic modeling



Topic Modeling | BERTopic



Topic Modeling | BERTopic

Opportunities

- Flexibility: Easy to adapt individual modules
- Supports dynamic, hierachical, multimodal, multilingual modeling
- Benefits from the continous improvement LLM embeddings
- No need to predefine the number of clusters (with HDBSCAN)

Caveats

- Each document is assigned only one topic
- Can generate many outliers (with HDBSCAN)
- Disconnection between clustering and topic representation steps
- Flexibility can be exploited for misrepresentation

References

[1] Mikolov, 2013: Efficient Estimation of Word Representations in Vector Space
[2] Devlin, 2018: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
[3] Reimers, 2019: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
[4] Su, 2023: One Embedder, Any Task: Instruction-Finetuned Text Embeddings
[5] Muennighoff, 2023: MTEB: Massive Text Embedding Benchmark
[6] Schramowski, 2022: Large pre-trained language models contain human-like biases of what is right and wrong to do
[7] Dalvi, 2022: Discovering Latent Concepts Learned in BERT
[8] Belinkov, 2022: Probing classifiers: Promises, shortcomings, and advances
[9] Greenwald, 1998: Measuring Individual Differences in Implicit Cognition: The Implicit Association Test
[10] Caliskan, 2017: Semantics derived automatically from language corpora contain human-like biases

References

[11] Guo, 2021: Contextualized Word Embeddings Contain a Distribution of Human-like Biases
[12] Osgood, 1952: The nature and measurement of meaning.
[13] Mathew, 2020: The polar framework: Polar opposites enable interpretability of pre-trained word embeddings
[14] Engler, 2023: SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings
[15] Fraser, 2021: Understanding and countering stereotypes: A computational approach to the stereotype content model
[16] Ethayarajh, 2019: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings
[17] Blei, 2003: Latent Dirichlet Allocation
[18] Wu, 2024: A Survey on Neural Topic Models: Methods, Applications, and Challenges
[19] Miao, 2019: Discovering Discrete Latent Topics with Neural Variational Inference
[20] Grootendorst, 2022: BERTopic: Neural topic modeling with a class-based TF-IDF procedure

Study Approach

Minimal

- work with the slides

Standard

- minimal approach + check out BERTopic Best Practices: https://maartengr.github.io/BERTopic/getting_started/best_practices/best_practices.html

In-Depth

- standard approach + skim references 4, 6, 7

See you next time!