Text as Data: Embeddings

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

So far: we have been learning representations of the data

- Dictionary methods: document is represented as a count over the lexicon
- N-grams: document is a count over a vocabulary of phrases
- Text regressions: produce $\hat{\boldsymbol{y}}_i = f(\boldsymbol{x}_i; \hat{\theta})$ – a prediction for each document i
- Topic models: document is a vector of shares over topics

Limitations of bag-of-words representations

- Until now, $\boldsymbol{x}_i$ has been a "bag-of-words" representation.
- Bag-of-words representations disregard **syntax**
  - *"The terrorists killed American soldiers."* versus *"The American soldiers killed terrorists."*

    $\rightarrow$ These two sentences have the same bag-of-words representation
- Bag-of-words representations disregard **semantic proximity** between words
  - *"hi"* and *"hello"* are completely distinct features for predicting whether a message is greeting somebody
  - *"economics"* and *"sociology"* are distinct features for predicting whether a message is about the social sciences
- This class: Can we estimate text features that capture semantic proximity?

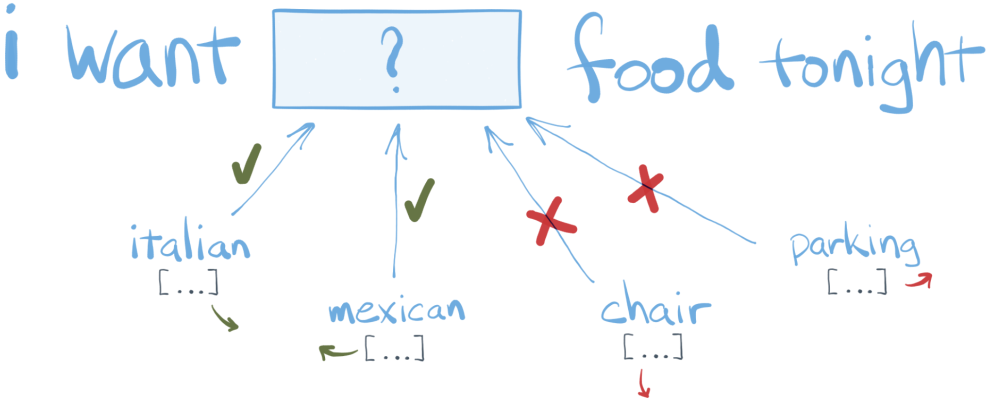
An example to build some intuition

Figure: Can you complete this text snippet?

Source: Patrick Harrison, S&P Global Market Intelligence

An Example to Build Some Intuition

Figure: Can you complete this text snippet?



Source: Patrick Harrison, S&P Global Market Intelligence

Language in context (and vice-versa)

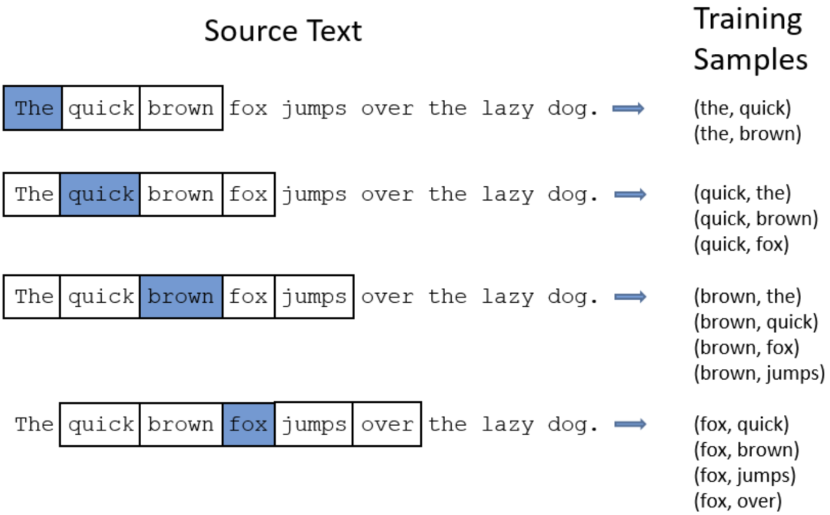
"You shall know a word by the company it keeps." (J. R. Firth, 1957)

- Neighboring words provide us with additional information to interpret a word's meaning
- In other words, **word co-occurrences capture context**
- This information is useful for machine learning applications
  - For example, document classification, machine translation, syntax prediction, machine comprehension, etc.

The brute force approach

- **Build a large word co-occurrence matrix** C
- Notations:
    - V is a vocabulary of $|V|$ words
    - M is an integer called the **window**
    - The M words preceding and the M words following a word constitute its **context**
- The cell (i, j) of C represents how many times the word i co-occurs with word j in the window.
- Each of the lines of C is a vector representation of a word that contains more information than one-hot vectors (i.e., bag-of-words).

Example for the window size



Source: Julian Gilyadov. Window size $M = 2$.

The limits to the brute force approach

- However, the resulting co-occurrence matrix C is **high-dimensional and sparse**
- As the vocabulary size increases, working with this matrix becomes intractable
- **Can we approximate C in a low-dimensional, dense vector space?** (i.e., such that $p << |V|$)

  $\rightarrow$ This is precisely what text embeddings are about

The first generation of embeddings

- The three most famous models are:
  - `Word2Vec`[1]
  - `GloVe`[2]
- We will look at `Word2Vec` in more detail

Tomas Mikolov

FOLLOW

Senior Researcher, CIIRC CTU
Verified email at cvut.cz

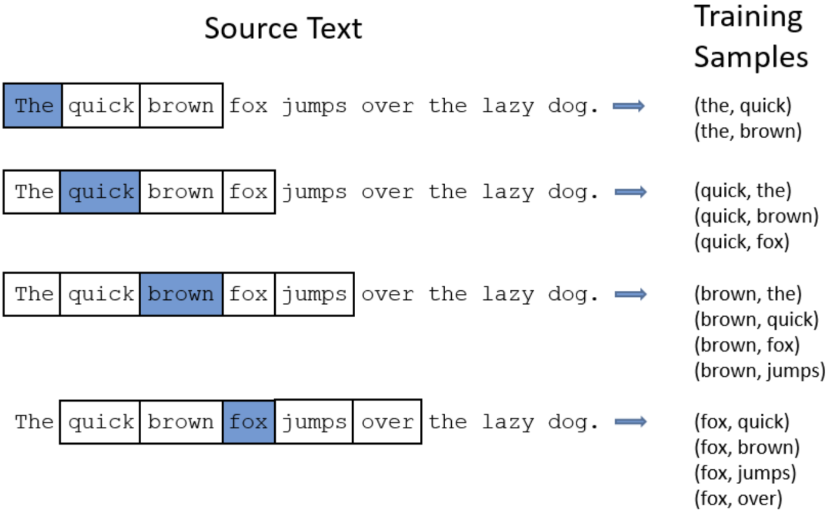
Artificial Intelligence Machine Learning Language Modeling Natural Language Processing

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	34060	2013

A “self-supervised” learning problem

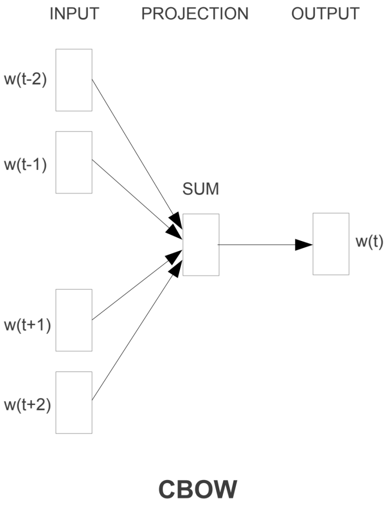
- `Word2Vec` reformulates learning word co-occurrences as two prediction tasks:
  - **Continuous Bag of Words (CBOW):** Given its context words, predict a focus word
  - **Skipgram:** Given a focus word, predict all its context words
- In both cases, the model results in a low-dimensional, dense vector space representation of C

Recall our example



Source: Julian Gilyadov. Window size $M = 2$.

CBOW: intuition



CBOW: likelihood

- Recall M, the size of the context window (often between 5 and 10)
- Given a sequence of T words, the log-likelihood is

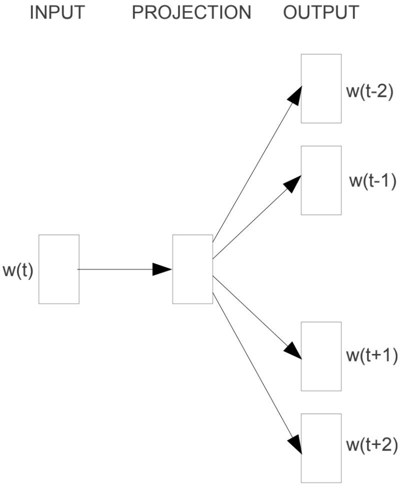
$$\frac{1}{T}\sum_{t=1}^{T}\log\left(P(w_t|\{w_{t+j}\}_{-M\leq j\leq M, j\neq 0})\right)$$

- The probability of observing the focus word w_t given its context words is

$$P(w_t|\{w_{t+j}\}_{-M\leq j\leq M, j\neq 0}) = \frac{\exp(w'_t\cdot\bar{u}_t)}{\sum_{k=1}^{|V|}\exp(w'_k\cdot\bar{u}_t)},$$

where $\bar{u}_t$ is the average of the context vectors for words in the context window, and w vectors are word vectors.

Skipgram – intuition



Skip-gram

Skipgram – likelihood

- Recall M, the size of the context window (often between 5 and 10)
- Given a sequence of T words, the log-likelihood is

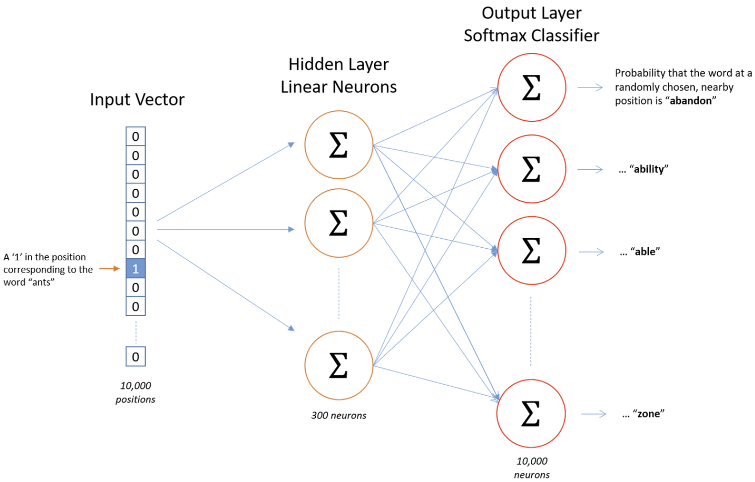
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-M\leq j\leq M, j\neq 0}\log\left(P(w_{t+j}|w_t)\right)$$

- The probability of observing context word w_{t+j} given the focus word w_t is

$$P(w_{t+j}|w_t) = \frac{\exp(y'_{t+j}\cdot w_t)}{\sum_{k=1}^{|V|}\exp(y'_k\cdot w_t)},$$

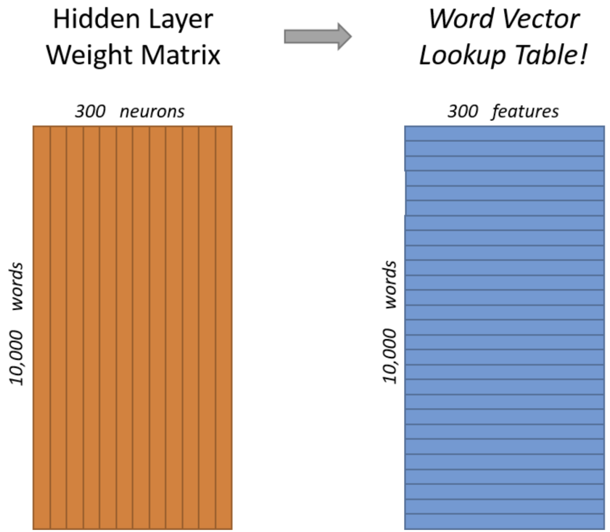
where y vectors are context vectors and w vectors are word vectors.

Neural network representation



Source: Julian Gilyadov. Contrary to most supervised learning tasks, the hidden layer is what we actually care about here. It represents the word vectors!

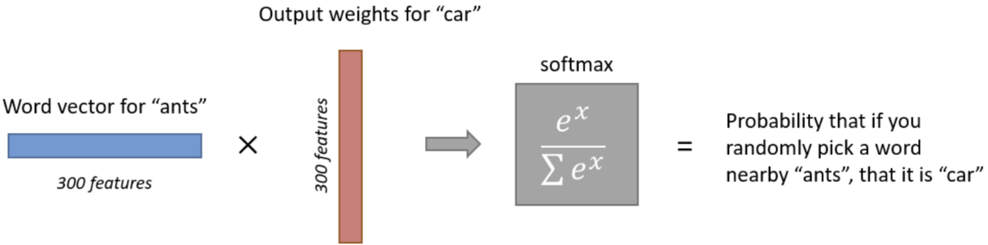
Lookup table



Source: Julian Gilyadov

$$
\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 12 & 19 \end{bmatrix}
$$

Source: Julian Gilyadov



Source: Julian Gilyadov

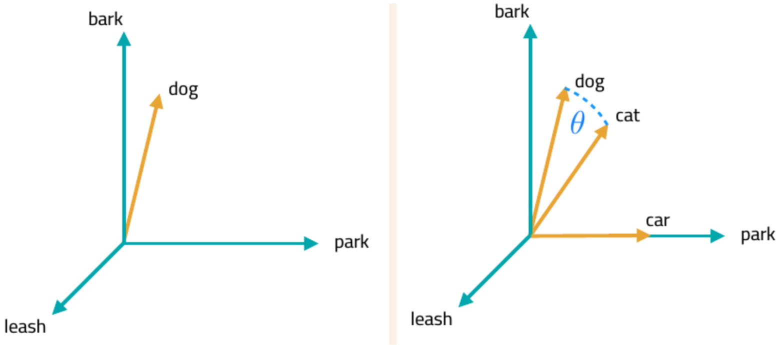
Distance between texts

- With embeddings, we can use linear algebra to understand **relationships between words**
- In particular, words that are geometrically close to each other are **similar**
- The standard metric for comparing vectors is **cosine similarity**:

$$\cos\theta = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

- When vectors are normalized, cosine similarity is:
  - Simply the dot product of both vectors
  - Proportional to the Euclidean distance (so you can use it, too)

Distance between texts



Visualizing embeddings

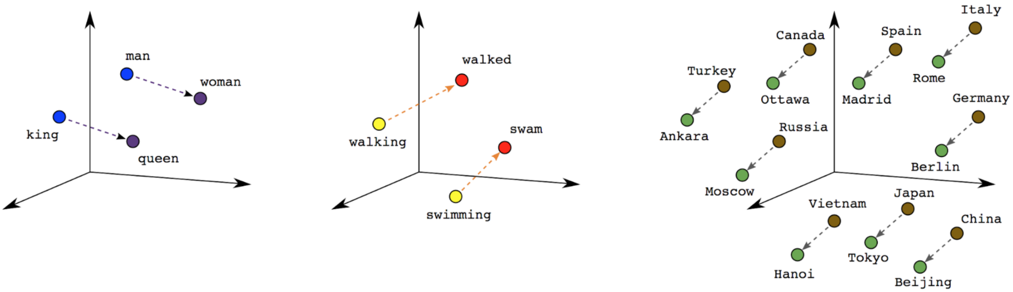
- One can also visualize the resulting embedding space by **projecting it on a two-dimensional space**
- Three commonly used techniques are:
    - Principal Component Analysis (PCA)
    - t-distributed stochastic neighbor embedding (t-SNE)
    - Uniform Manifold Approximation and Projection (UMAP)

Visualizing embeddings

Source: Ash et al. 2024

Basic arithmetic often carries meaning

- `Word2vec` algebra can depict conceptual, analogical relationships between words.
- *e.g.*, $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\textit{queen}}$



Male-Female Verb Tense Country-Capital

Some refinements

- The main assumption behind `word2vec` is that **context words are exchangeable**
- In other words, the ordering of words is not accounted for
- Recent models relax this assumption; they are called **sequence models**...
- .. and consistently outperform previous language models in various tasks

Pros and Cons

- **Pros**
  - Many pre-trained models for different languages are freely available online
  - Many packages to train models from scratch or fine-tune existing models to a specific corpus
  - Often, they provide sizable gains in prediction accuracy
- **Cons**
  - Clear loss of interpretability relative to bag-of-words
  - Neighbouring words are not the only forms of context (e.g., metadata)

References I

Ash, Elliott, Germain Gauthier, and Philine Widmer (2024). "Relatio: Text semantics capture political and economic narratives". In: *Political Analysis* 32.1, pp. 115–132.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.