Natural Language Processing

Dr. Manuel Pita & Zuil Pirola

UNIVERSIDADE
LUSÓFONA

Semantic

Semantics is the **study of meaning** in language, focusing on how words, phrases, and sentences convey and interpret meaning. It explores relationships like synonymy, ambiguity, and context, with applications in linguistics, philosophy, and **computer science**.

Semantic

- **Lemmas**: representing its canonical form used for grouping related words. For example, "dorminhoco", "dormindo", and “**dormir**”
- **Polysemy** refers to the phenomenon where a single word or expression has multiple related meanings ("banco" and “manga")
- **Synonym** is a word or phrase that has the same or nearly the same meaning as another word in the same language. Synonyms are used to avoid repetition or to add variety and precision to language. For example: Water = H2O and House = Home

Semantic

Cats and **dogs** are **not similar** in nature, but both are **pets**, making them more semantically related to each other than to the word “**sky**."

Topic models, such as Latent Dirichlet Allocation (**LDA**), can capture these relationships by identifying patterns where related **words frequently appear together** in textual data.

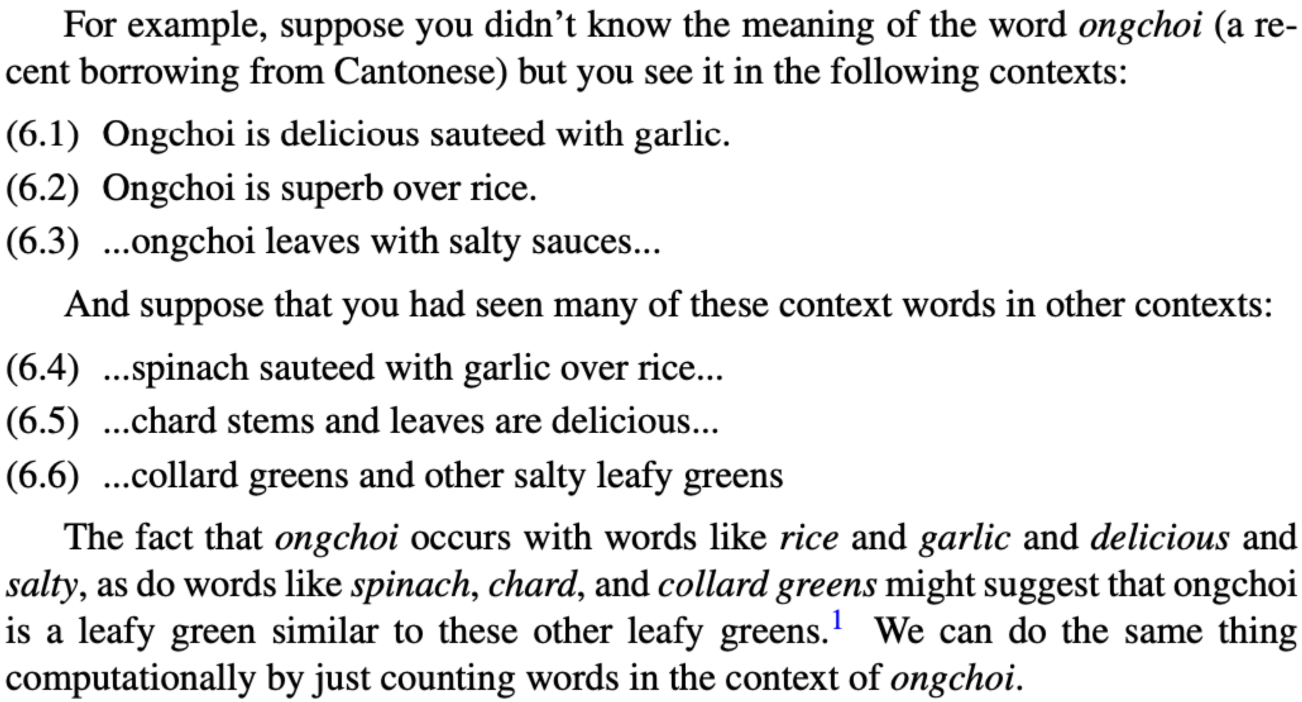
Meaning **beyond their literal sense**. For example, "love" has a positive **connotation**, while "hate" carries a negative one. These connotations are essential in **sentiment analysis**

Vector Semantics

Vector Semantic

In NLP, **vector semantics** defines the meaning of a word based on its **distribution in language use**, i.e., its neighboring words or grammatical context. This approach relies on the idea that words occurring in similar contexts have similar meanings. Words are represented as vectors in a high-dimensional space, where words with similar contexts (e.g., "dog" and "cat") are placed closer together. This concept underpins techniques like **word embeddings** (e.g., Word2Vec), allowing for tasks like measuring word similarity and semantic analysis.

Vector Semantic



For example, suppose you didn't know the meaning of the word *ongchoi* (a recent borrowing from Cantonese) but you see it in the following contexts:

(6.1) Ongchoi is delicious sauteed with garlic.

(6.2) Ongchoi is superb over rice.

(6.3) ...ongchoi leaves with salty sauces...

And suppose that you had seen many of these context words in other contexts:

(6.4) ...spinach sauteed with garlic over rice...

(6.5) ...chard stems and leaves are delicious...

(6.6) ...collard greens and other salty leafy greens

The fact that *ongchoi* occurs with words like *rice* and *garlic* and *delicious* and *salty*, as do words like *spinach*, *chard*, and *collard greens* might suggest that ongchoi is a leafy green similar to these other leafy greens.[1] We can do the same thing computationally by just counting words in the context of *ongchoi*.

Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

Vector Semantic

Vector semantics represents **words as points in a multidimensional semantic space**, derived from the **distribution of word neighbors' embeddings**. These word representations are called embeddings, though the term is sometimes specifically used for dense vectors like **word2vec**.

Vector Semantic

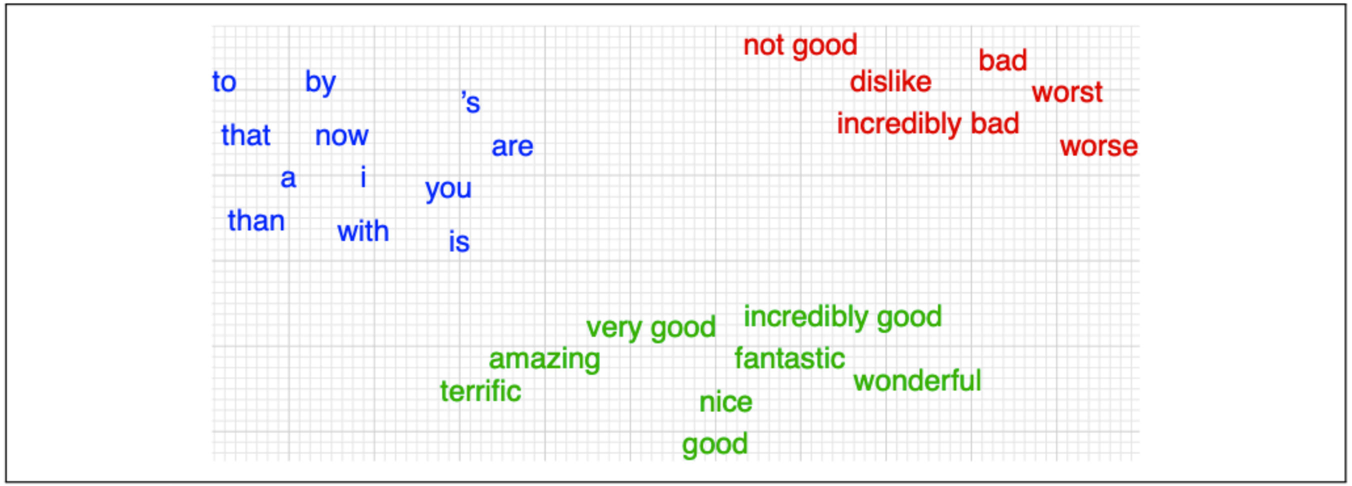


Figure 6.1 A two-dimensional (t-SNE) projection of embeddings for some words and phrases, showing that words with similar meanings are nearby in space. The original 60-dimensional embeddings were trained for sentiment analysis. Simplified from Li et al. (2015) with colors added for explanation.

Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

Vector Semantic

Similar words have similar vectors as they **appear in similar documents**. The term-document matrix represents a word's meaning through the documents it frequently occurs in.

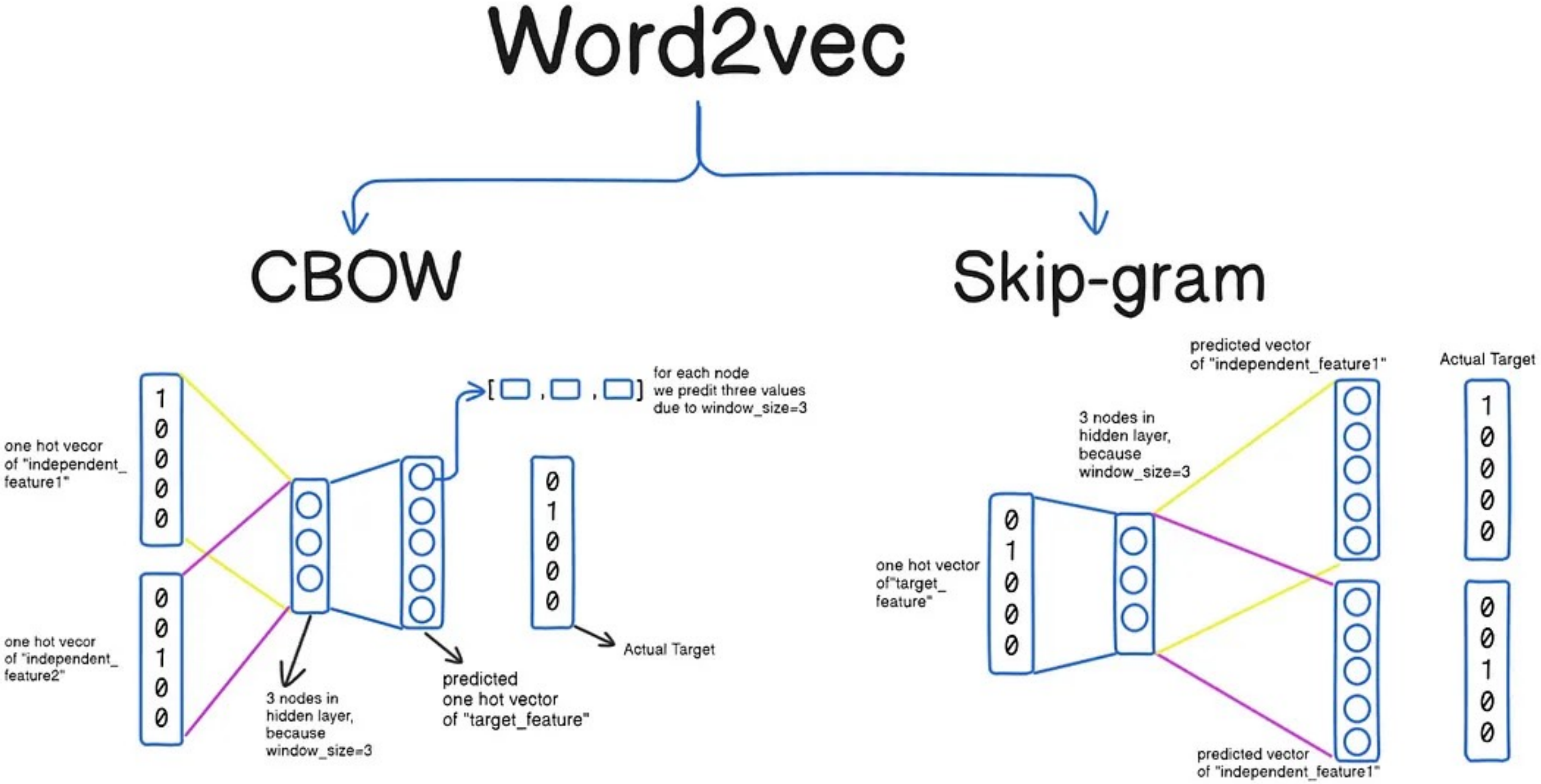
Word-word matrix, also known as the term-context matrix, where **columns are labeled by words instead of documents**.

Word2vec

Word2Vec

Word2vec is a method for generating **word embeddings**, where each word in the vocabulary is represented by a **fixed, static vector**. These embeddings are learned through a **neural network**, capturing semantic relationships between words based on their contexts in large text corpora. Word2vec creates one unique representation for each word, and these embeddings remain constant once trained, meaning they do not change dynamically during usage.

Word2Vec


Word2vec
CBOW
Skip-gram
one hot vecor of "independent_ feature1"
one hot vecor of "independent_ feature2"
1 0 0 0 0
0 0 1 0 0
3 nodes in hidden layer, because window_size=3
predicted one hot vector of "target_feature"
for each node we predit three values due to window_size=3
0 1 0 0 0
Actual Target
predicted vector of "independent_feature1"
Actual Target
one hot vector of"target_ feature"
0 1 0 0 0
3 nodes in hidden layer, because window_size=3
1 0 0 0 0
0 0 1 0 0
predicted vector of "independent_feature1"

https://medium.com/@fraidoonomarzai99/word2vec-cbow-skip-gram-in-depth-88d9cc340a50

Word2Vec

Key concepts:

- Window
- Target
- Size of vector
- One hot vector
- Logistic Regression

Word2Vec

CBOW:

- **Pros:** Faster, efficient for frequent words, works well with large corpora.
- **Cons:** Less effective for rare words, loses complex context details.

Skipgram:

- **Pros:** Better for rare words, captures complex semantic relationships.
- **Cons:** Slower, more computationally expensive.

Summary:
CBOW is faster and more efficient, while Skipgram excels with rare words and complex context.

Bias and embeddings problems

- Static Vectors, Ambiguity and Polysemy
- Bias in Embeddings
- Impact on Search and Automated Systems
- Training Data Dependency
- Sentiment and Emotional Nuances (bom e ótimo)
- Poor Generalization (unseen words)