NLP

GloVe: Global Vectors

Natural Language Processing

Agenda

- Semantic Analysis and its importance
- The Co-occurrence Matrix
- GloVe
- GloVe vs Word2vec

What is Semantic Analysis? Why is it important?

- **Semantic analysis** is a branch of **natural language processing** and **machine learning** that aids in comprehending the context of any text and the emotions that might be expressed in a sentence. Machine translation, chatbots, search engines, and text analytics all make use of semantic analysis.

- Why is **Semantic Analysis** important?

    - Semantic Analysis helps to provide the text data more context so that the model can understand it more accurately for a given NLP task.

    - For example, elements like Hyponymy, Polysemy, Synonymy, etc is essential for question-answer systems like chatbots to gather insight from textual data and extract information.

The Co-occurrence Matrix

- **Global Vector (GloVe)** models word vectors using the computed statistics of the **co-occurrence matrix**.

- A co-occurrence matrix is an **NxN matrix** where each row and each column represents a **unique word** in a given **corpus**.

	I	enjoy	deep	nlp	learn	fly	like	.
I	0	1	0	0	0	0	2	0
enjoy	1	0	0	0	0	1	0	0
deep	0	0	0	0	1	0	1	0
nlp	0	0	0	0	0	0	1	1
learn	0	0	1	0	0	0	0	1
fly	0	1	0	0	0	0	0	1
like	2	0	1	1	0	0	0	0
.	0	0	0	1	1	1	0	0

The Co-occurrence Matrix

- Each entry in the co-occurrence matrix represents the **number of times** the word **'i'** has occurred with the word **'j'**.

	I	enjoy	deep	nlp	learn	fly	like	.
I	0	1	0	0	0	0	2	0
enjoy	1	0	0	0	0	1	0	0
deep	0	0	0	0	1	0	1	0
nlp	0	0	0	0	0	0	1	1
learn	0	0	1	0	0	0	0	1
fly	0	1	0	0	0	0	0	1
like	2	0	1	1	0	0	0	0
.	0	0	0	1	1	1	0	0

The Co-occurrence Matrix

- The relationship of the words can be studied by observing the ratio of their **co-occurrence probabilities**.

- Let **P(k|ice)** be the **probability of observing the word 'k' with 'ice'**, and **P(k|steam)** be the **probability of observing the word 'k' with the word steam**.

	k = solid	k = gas	k = water	k = fashion (random)
P(k|ice)	high	low	high	low
P(k|steam)	low	high	high	low
P(k|ice)/ P(k|steam)	>1	<1	~1	~1

Co-occurrence Matrix

- As one might assume, **ice co-occurs with solids** more frequently **than with gases**, but **steam co-occurs with gases** more frequently **than with solids**. Both words usually occur with their **shared property water**, and both occur infrequently with the **unrelated word fashion**. **Large values** (much greater than 1) correlate well enough with **ice-specific features**, while **small values** (much less than 1) correlate **strongly with steam-specific qualities**.

- In this way, **the probability ratio encodes some basic meaning connected to the general idea of thermodynamic phase.**

	k = solid	k = gas	k = water	k = fashion (random)
P(k|ice)	high	low	high	low
P(k|steam)	low	high	high	low
P(k|ice)/ P(k|steam)	>1	<1	~1	~1

GloVe

- **GloVe** attempts to generate word vectors by **using the global co-occurrence relationship.** GloVe's training objective is to learn **word vectors whose dot product equals the logarithm of the probability of occurrence of the words.**

- In the previous slide, we saw how **word-to-word co-occurrence probabilities have the potential to encode some form of meaning** by looking at an example where we captured certain information about the thermodynamic phase of matter using probability ratios.

- As these word-to-word co-occurrence probabilities have the potential to determine meaning, t**hey are also encoded in the vector representation**.

- Therefore, the **embeddings** created using **GloVe** capture a significant amount of **semantic** and **syntactic** meaning.

Word2vec VS GloVe

- **Both models** differ in the way they are trained, and hence they **output different word vectors**.

- **The GloVe** model is based on **global word to word co-occurrence counts** taking the whole corpus into consideration, whereas **Word2vec** uses **co-occurrences of local context (neighbouring words)**.

- **GloVe** learns embeddings by **constructing the co-occurrence matrix** - on the other hand the **Word2vec** model learns by making predictions by **taking context words as inputs and predicting the target words**.

- **GloVe** is thought to capture the **semantic** as well as the **syntactic** meaning of the words in its embeddings **better than Word2vec**.

Summary

So, in order to summarize:

- **Semantic analysis** is a branch of Natural Language processing and machine learning that **aids in comprehending the context of any text and the emotions that might be expressed in a sentence**.

- **GloVe** learns the embeddings **by constructing the co-occurrence matrix**, on the other hand the **Word2vec** model learns making prediction **by taking context words as inputs** and **predicting target words**.

- **The embeddings** created **using GloVe capture** as much **semantic and syntactic meaning** as possible.

Great Learning

Happy Learning !