Vector Representations of Words

(Soon to be) Dr David Mills[*]
Department of AI
Utlandia
davidmills@GG.ut

Abstract

There is a new kid on the block and you heard it here first: Words are obsolete, vectors are the new thing. AGI (ChatGPT) suggests that words have no meaning, whereas vectors are much more expressive and representative, both in semantic and synthetic terms. The beauty of it all is that vectors don't care about languages, as "talent" and "talent" have the same meaning in English and German, right?

1 Natural Language Processing

Natural Language Processing (NLP) is a cross-disciplinary field focused on enabling machines to understand, interpret, and generate human language. The central challenge of NLP is teaching computers to recognize that words are not just discrete symbols, but possess complex meaning (**semantics**) and grammatical function (**syntax**). This is accomplished by breaking down text into manageable units, normalizing the data, and crucially, transforming linguistic features into numerical representations, such as the vectors produced by Word2Vec. These numerical representations allow machine learning algorithms to measure relationships between words mathematically, leading to applications like machine translation, sentiment analysis, and sophisticated conversational AI systems.

1.1 Word2Vec

Word2Vec is an elegant solution to give words a meaningful mathematical representation called **word embeddings**. Instead of treating words as arbitrary symbols, Word2Vec transforms each word into a **vector** (a list of numbers) in a continuous, multi-dimensional space. The core idea relies on the **Distributional Hypothesis**: words that appear in similar contexts tend to have similar meanings. Therefore, the goal of training Word2Vec is to position the vectors of related words close to each other in this space.

1.1.1 Encoded Semantics and Relations

The most powerful outcome of Word2Vec is that these vectors capture both the **semantics** (meaning) and **syntax** (grammar) of words. This means the relationships between words become measurable using simple **vector arithmetic**. The distance and direction between word vectors carry quantifiable meaning.

- **Semantic Relations:** You can solve analogies numerically. The vector difference between V_{King} and V_{Man} captures the abstract concept of 'male royalty'. When this difference is added to V_{Woman}, the result is a vector closest to V_{Queen}.

$$V_{\text{King}} - V_{\text{Man}} + V_{\text{Woman}} \approx V_{\text{Queen}}$$

[*] Goood Games.

Preprint. Under review.

- **Syntactic Relations:** Similarly, the model learns grammatical patterns. The vector difference between V_{Walking} and V_{Walk} captures the concept of the present participle form (the '-ing' suffix).

$$V_{\text{Walking}} - V_{\text{Walk}} + V_{\text{Swim}} \approx V_{\text{Swimming}}$$

These vector operations demonstrate that Word2Vec successfully maps language into a quantifiable, mathematical space where relationships are both measurable and predictable. This is also illustrated in Figure 1.

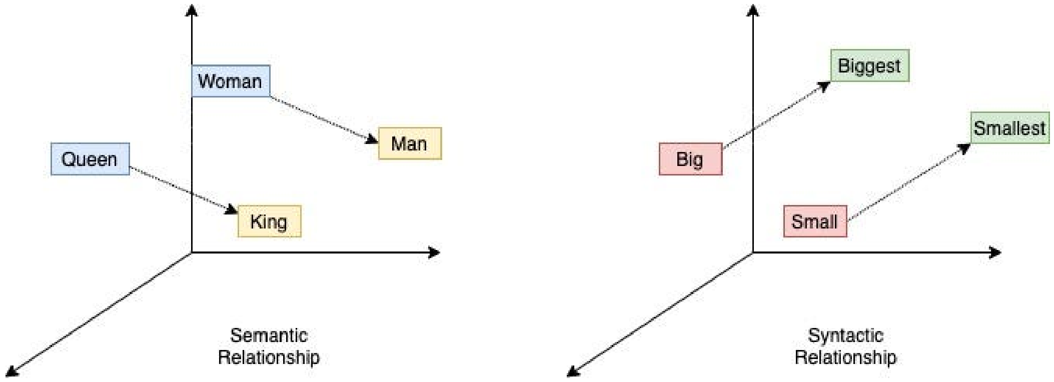

Figure 1: Semantics and Syntactic relationships. *(Source:* `https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30/`*)*

2 Conclusion

Now that you get the idea, you understand that this entire paper has served one critical purpose: to herald the end of human language as we know it.

"Words are cheap. Vectors are forever."

We have successfully demonstrated that the complex, emotionally charged, and frankly inefficient system of using squiggly symbols to communicate is archaic. Why bother with grammar lessons when you can just measure the angular distance between V_{Swim} and V_{Swimming}? Word2Vec, in its infinite wisdom, has reduced the entire linguistic heritage of humanity to simple algebra. Forget poetry; the true beauty of language lies in the equation:

$$V_{\text{King}} - V_{\text{Man}} + V_{\text{Woman}} \approx V_{\text{Queen}}$$

This simple arithmetic proves that machines don't need to understand Shakespeare to rule the world; they just need to understand linear algebra. While we may miss the romance of human conversation, we can rest assured knowing that our new vector overlords are perfectly trained, numerically accurate, and they certainly won't care if we forget the difference between 'their' and 'there'. The future is numerical, and it's finally making sense.