CL Seminar 2

Word Embeddings as Tuples of Feature Probabilities
by Siddharth Bhat, Alok Debnath, Souvik Bannerjee and Manish Shrivastava

Abhinav S Menon

Overview

- Introduction
- Fuzzy Sets
- Representation and Operations
- Similarity
- Analogy
- Experiments and Results

Introduction

Word embeddings are mechanism to represent words as vectors. Typically, a word vector is supposed to be physically close to the vectors of words to which it is close in meaning.

The actual values of a word vector depend on the context in which it occurs.

In this paper, the authors define a new mechanism of representing words using the mathematical notions of fuzzy sets.

"Can a single word representation mechanism account for lexical similarity and analogy, compositionality, and lexical entailment **and** *be used to detect and resolve polysemy?"*

Fuzzy Sets

A fuzzy set is defined as a set of pairs $\{(x, \mu_A(x)), x \in \Omega\}$, where $\mu_A(x)$ represents the possibility of x belonging to A and Ω is the universal set.

In other words, a fuzzy set is a set with probabilistic set membership.

Fuzzy Sets as Tuples

Suppose the universe Ω has cardinality n. Let $\Omega = \{x_1, x_2, \ldots, x_n\}$. Then we can represent any fuzzy set A as an n-tuple A'.

The i^{th} entry of A' represents the probability of the i^{th} element of Ω belonging to A, *i.e.*, $A'_i = \mu_A(x_i)$.

Operations on Fuzzy Sets

We define seven operations on fuzzy sets:

$$(A \cap B)_i = A_i \times B_i \text{ (set intersection)}$$

$$(A \cup B)_i = A_i + B_i - A_i \times B_i \text{ (set union)}$$

$$(A \sqcup B)_i = max(1, min(0, A_i + B_i)) \text{ (disjoint set union)}$$

$$(\neg A)_i = 1 - A_i \text{ (set complement)}$$

$$(A \setminus B)_i = A_i - \min A_i, B_i \text{ (set difference)}$$

$$(A \subseteq B)_i = \forall x \in \Omega : A_i \leq B_i \text{ (set inclusion)}$$

$$|A| = \sum_{i \in \Omega} A_i \text{ (cardinality)}$$

Entropy

Entropy (also known as Shannon entropy) is a measure of the "information" contained in the outcome of a random variable.

For a single variable x with probabilities $p_1, p_2, \ldots, p_n$, the entropy $H(x)$ is defined as

$$\sum_{i=1}^{n} p_i \ln\left(\frac{1}{p_i}\right).$$

This notion is extended to fuzzy sets by summing over the entropies of its members.

$$H(A) = \sum_{x_i \in \Omega} A_i \ln\left(\frac{1}{A_i}\right) + (1 - A[i]) \ln\left(\frac{1}{1 - A_i}\right)$$

Representation

The idea presented in the paper is to represent words as fuzzy sets (or tuples of probabilities) instead of simply vectors.

In this representation, each dimension is associated with a feature. The component of the vector along that dimension is the probability of the word having that feature.

Thus, these representations are referred to as “tuples of feature probabilities”.

Representation

Ordinary word vectors are converted to tuples of feature probabilities by exponentiating and normalising their components across the entire vocabulary. For example, given a word vector v in vocabulary V,

$$\forall i : \hat{v}_i = \frac{\exp v_i}{\sum_{w \in V} \exp w_i},$$

where $\hat{v}$ is the tuple of feature probabilities for the given word.

This tuple is a fuzzy set, and its entries are interpreted as follows: if the i^{th} dimension is a property, then $\hat{v}_i$ is the probability that v has that property.

Operations

The fuzzy set operations all have a semantic interpretation.

The operation	gives us
$w_1 \cap w_2$	a word which has features common between w_1 and w_2.
$w_1 \cup w_2$	a word which has the features of both w_1 and w_2.
$w_1 \setminus w_2$	a word which has the features of w_1 but not w_2.

Operations

The logical relation $w_1 \subseteq w_2$ (feature inclusion) is equivalent to w_1 being a hyponym of w_2.

The entropy of w_1 indicates the level of certainty to which w_1 possesses the features in its representation. Thus, a word having a high entropy value are mostly those that have some probability of possessing several features.

This allows us to make use of entropy to find function words in a corpus.

Similarity

Similarity is central to the idea of word vectors; it is one of the most important relations a representation must be able to formalise.

One notion of similarity provided by tuples of feature probabilities is feature difference (or fuzzy set difference).

Further, definitions of similarity between probability distributions (like K-L divergence and cross-entropy) can be used, which are also inherently asymmetric.

Similarity

K-L Divergence

In an information theoretic sense, it is the number of extra bits needed to store P under the (false) assumption that it follows Q.

$$D(P||Q) = \sum_{i=1}^{n} P_i \ln \left(\frac{P_i}{Q_i} \right)$$

Cross-Entropy

Cross-entropy is closely related to K-L divergence, but it takes into account the entropy of P as well.

$$H(P, Q) = H(P) + D(P||Q)$$

Analogy

In the case of ordinary word vectors, analogy is constructed using vector addition and subtraction, *i.e.*, if we know $(a : b)$ and x, and we need to find y such that $a : b :: x : y$, we calculate it as $y = (b + x) - a$.

We use a closely related formula for the corresponding feature probability tuples – $y = (b \cup x) \setminus a$.

Note the use of non-disjoint set union in this formula.

Experiments and Results

Similarity and Analogy

For similarity, the fuzzy representation performs consistently better than ordinary representations using both K-L divergence and cross-entropy. Further, cross-entropy usually outperforms K-L divergence.

In the case of analogy, the fuzzy representation does better than ordinary representations at lower dimensions (50 or 100).

Function Word Detection

The entropy method detects function words correctly much more consistently than the frequency method.

Experiments and Results

Compositionality

Two fuzzy vectors can be combined to represent the compound formed from the corresponding words. This operation represents composition better than the ordinary representation at smaller dimensions.

Dimensionality Analysis and Feature and Representations

This representation does not appear to scale well as dimensions increase.

This is probably because the representation relies on the probability distribution across the entire vocabulary, making it sparse when the number of dimensions increases.

References

Bhat, Siddharth, et al. “Word Embeddings as Tuples of Feature Probabilities.” Proceedings of the 5th Workshop on Representation Learning for NLP. 2020.