

Universitat Politècnica de Catalunya
MAMMEE
Final Thesis presentation

An analysis of word embedding spaces and regularities

Manuel Gijón Agudo

Index

- 1 Introduction
- 2 Resources
- 3 Metodology and evaluation
- 4 Results
- 5 Conclusions

Introduction

Word embeddings (WE) used to come from training a Neuronal Network with a corpus of text.

Introduction

Word embeddings (WE) used to come from training a Neuronal Network with a corpus of text.

WE are widely use due to their ability to capture semantic relationships between words as relations between vectors in high dimensional spaces in several applications such that machine translation (similarity between words) or image annotation.

Goals

- Establish a criterion to evaluate WE based on how strong the semantic relations are embodied in the distances between pairs of words.
- Use that criteria to study the performance of different distances, some of them well known.
- Point out the different behaviour of the two most used WE.
- Present a new and competitive distance.

- The GloVe model (Global Vectors for Word Representation) was developed and introduced in 2014 [1] by researchers of Stanford University.
- We will work with dimensions 50, 100, 200 and 300.
- The training corpus are Wikipedia 2014 and Gigaword-5.

- Word2Vec was created by Mikolov et al. [2], [3] and [4].
- The WE available trained with this methods has dimension 300.
- Two algorithms are described: Continuous Bag-of-Words (for predicting a target word given the words around it) and Skip-Gram (to predict the words surrounding a given word) [5], [6].

WordNet [7] a database. It is one of the most used resources in Natural Language Processing and Representation Learning.

WordNet [7] a database. It is one of the most used resources in Natural Language Processing and Representation Learning.

Synset

A synset is a set of terms that represent an unique idea or concept. All the words included in a synset are considered synonymous.

The English version of this database includes information over 117,000 different synsets and their semantic relations (hyponymy, hypernymy, etc.).

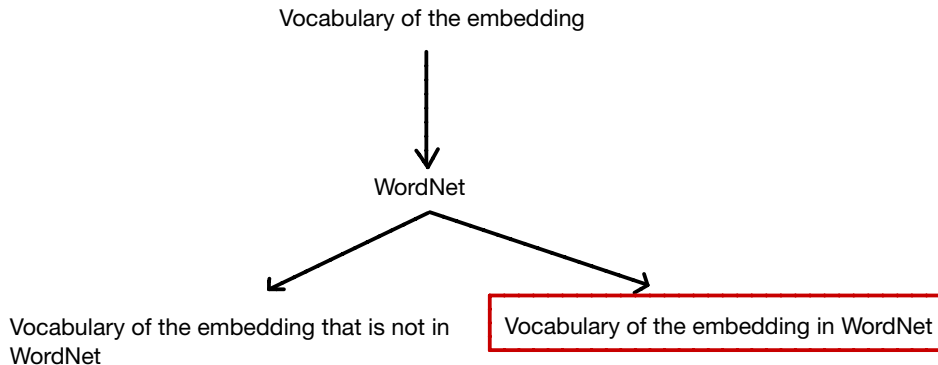
Synset: an example

Hamburger

- Hamburger (an inhabitant of Hamburg)
 - direct hypernym:
 - German (a person of German nationality)
 - sister term
 - German (a person of German nationality)
 - East German (a native/inhabitant of the former GDR)
 - Bavarian (a native/inhabitant of Bavaria)
 - derivationally related form
 - Hamburg (a port city in northern Germany on the Elbe River that was founded by Charlemagne in the...)

Figure: Information contained in the Synsets “Hamburguer”

WordNet in use: preparing the vocabulary



WordNet in use: defining relations

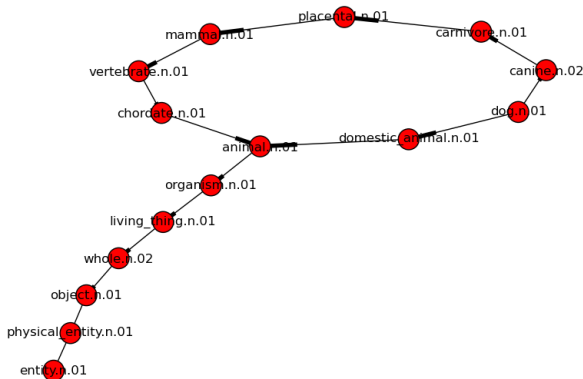


Figure: Example of relations between synsets.

Most used distances

- Euclidean:

$$d(u, v) = \left(\sum_{i=1}^N u_i - v_i \right)^{1/2}$$

- Cosine:

$$d(u, v) = 1 - \frac{u \cdot v}{||u|| ||v||}$$

Used distances

- Correlation:

$$d(u, v) = 1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{\|u - \bar{u}\| \|v - \bar{v}\|}$$

- Canberra:

$$d(u, v) = \sum_{i=1}^N \frac{|u_i - v_i|}{|u_i| + |v_i|}$$

- Braycurtis:

$$d(u, v) = \sum_{i=1}^N \frac{|u_i - v_i|}{|u_i + v_i|}$$

The curse of dimensionality

“In high dimensional spaces the concept of proximity, distance or nearest neighbour may not even be qualitatively meaningful”.

The curse of dimensionality

“In high dimensional spaces the concept of proximity, distance or nearest neighbour may not even be qualitatively meaningful”.

From [8] we know that concepts such as nearest neighbour or even distance loss their meaning for dimensions greater than 15. In [9] is proved that from the L_p family of distances, greater the p worst the effect of dimensionality.

The curse of dimensionality: an example

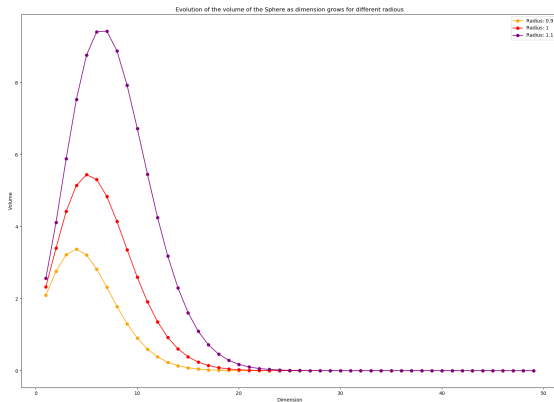


Figure: Volume of the sphere (dimension N) for radii 0.9, 1 and 1.1.

- 1 Filtering the vocabulary of the embedding using WordNet.
- 2 Select a relation between words and use WordNet to split the set of words.
- 3 We take distances between words in both sets obtaining two empirical distributions.
- 4 We use a statistical test to study the distributions.

Kolmogorov-Smirnov KS test

The KS test is the tool that we are going to use to compare our empirical distributions.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Kolmogorov-Smirnov KS test

The KS test is the tool that we are going to use to compare our empirical distributions.

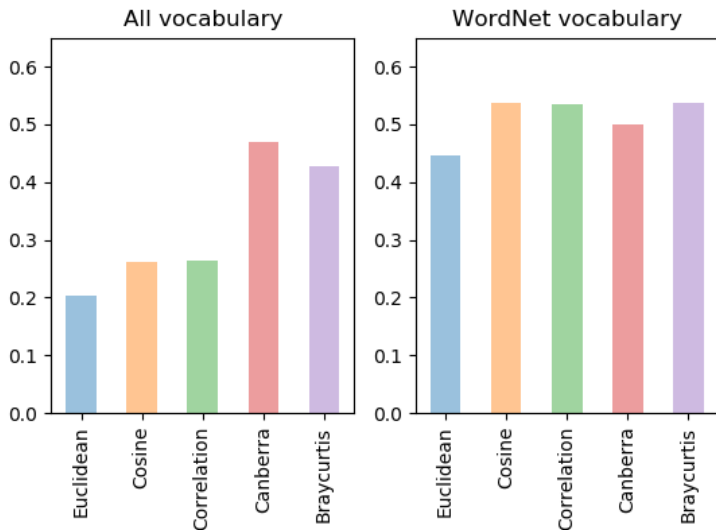
$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Null hypothesis

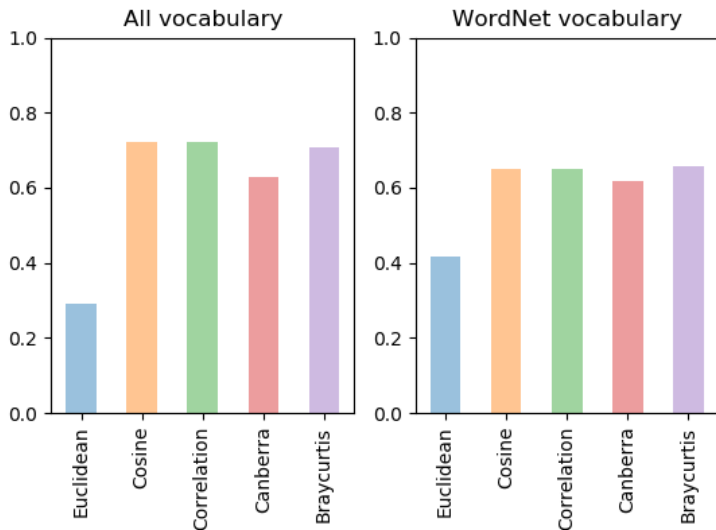
The two samples, $F_{1,n}$ and $F_{2,m}$, follows the same distribution at a level of significance α . We reject if:

$$D_{n,m} > \left(\sqrt{-\frac{1}{2} \ln \alpha} \right) \sqrt{\frac{n+m}{nm}}$$

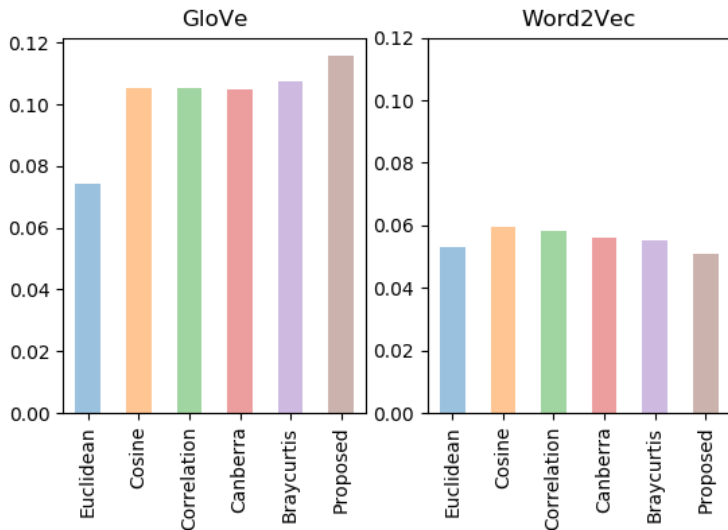
GloVe, separate synonyms



Word2Vec, separate synonyms



Separate synonyms and antonyms



Proposed distance

Let $u = (u_1, \dots, u_N)$ and $v = (v_1, \dots, v_N)$ vectors of \mathbb{R}^N .

Definition

$d(u, v) =$ “number of coordinates with different sign”.

For simplicity: consider same sign positive and positive and less or equal to zero and less or equal to zero.

Proposed distance: time performance

0.325

```
2019-07-13 10:58:55,801 - efficiency of proposed norm - INFO -  
-----  
2019-07-13 10:58:55,801 - efficiency of proposed norm - INFO - Starting with Euclidean  
2019-07-13 10:58:56,126 - efficiency of proposed norm - INFO - Euclidean finished  
2019-07-13 10:58:56,126 - efficiency of proposed norm - INFO -
```

```
2019-07-13 10:58:56,126 - efficiency of proposed norm - INFO - Starting with Proposed  
2019-07-13 10:58:56,202 - efficiency of proposed norm - INFO - Proposed finished  
2019-07-13 10:58:56,203 - efficiency of proposed norm - INFO -
```

0.076

0.338

```
2019-07-13 10:58:57,769 - efficiency of proposed norm - INFO -  
-----  
2019-07-13 10:58:57,769 - efficiency of proposed norm - INFO - Starting with Euclidean  
2019-07-13 10:58:58,107 - efficiency of proposed norm - INFO - Euclidean finished  
2019-07-13 10:58:58,107 - efficiency of proposed norm - INFO -
```

```
2019-07-13 10:58:58,107 - efficiency of proposed norm - INFO - Starting with Proposed  
2019-07-13 10:58:58,185 - efficiency of proposed norm - INFO - Proposed finished  
2019-07-13 10:58:58,185 - efficiency of proposed norm - INFO -
```

0.078

0.327

```
2019-07-13 10:58:59,407 - efficiency of proposed norm - INFO -  
-----  
2019-07-13 10:58:59,407 - efficiency of proposed norm - INFO - Starting with Euclidean  
2019-07-13 10:58:59,734 - efficiency of proposed norm - INFO - Euclidean finished  
2019-07-13 10:58:59,735 - efficiency of proposed norm - INFO -
```

```
2019-07-13 10:58:59,735 - efficiency of proposed norm - INFO - Starting with Proposed  
2019-07-13 10:58:59,811 - efficiency of proposed norm - INFO - Proposed finished  
2019-07-13 10:58:59,811 - efficiency of proposed norm - INFO -
```

0.076 25.3%

0.330

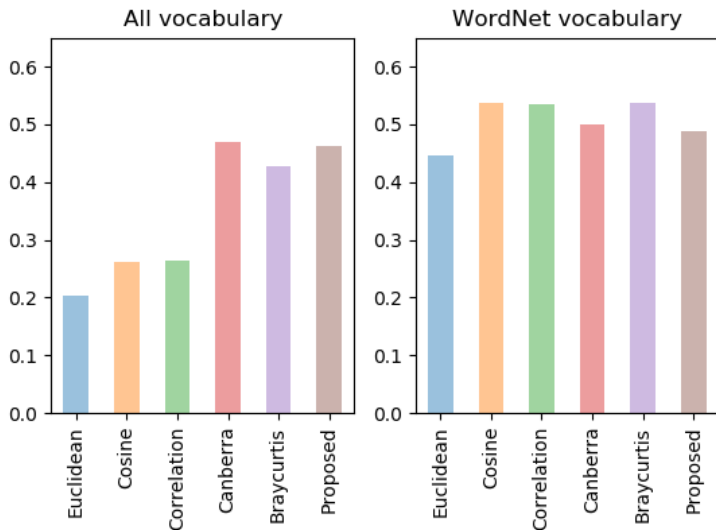
```
2019-07-13 10:59:00,831 - efficiency of proposed norm - INFO -
```

```
-----  
2019-07-13 10:59:00,831 - efficiency of proposed norm - INFO - Starting with Euclidean
```

0.0767

Figure: Time performance between Euclidean and Proposed distance for vectors of dimension 100000.

Proposed distance: consistency







Conclusions

- 1 We have establish a criterion to evaluate WE based on how strong the semantic relations are embodied in the distances between pairs of words.
- 2 Under this criteria, we know which ones of the distances performs better for every scenario.
- 3 We know that GloVe and Word2Vec present different behaviour.
- 4 We have introduced a new and competitive distance.

- Publication:
Analyzing distances in word embeddings and their relation with seme analysis. *Manuel Gijón Agudo, Armand Vilalta Arias and Dario García-Gasulla*
Accepted in Catalan Conference on Artificial Intelligence (CCIA-2019)
- We have created a python package called WER (Word Embedding Research) available in the web address <https://github.com/MGijon/WER>.

References

-  P. Jeffrey, R. Socher, and C. Manning. “Glove: Global vectors for word representation”. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (2014).
-  T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. “Distributed Representations of Words and Phrases and their Compositionality” (2013).
-  T. Mikolov, K. Chen, G. Corrado, J. Dean. “Efficient Estimation of Word Representations in Vector Space” (2013).
-  T. Mikolov, W. Yih, G. Zweig. “Linguistic Regularities in Continuous Space Word Representations” (2013).
-  Y. Goldberg and O. Levy. “Word2Vec Explained: Deriving Mikolov et al.’s Negative-Sampling Word-Embedding Method” (2014).