Semantic Similarity Detection

BTP Advisor: Dr. Amit Praseed.

Project Team: Hrishabh Pandey (064), Ayush Gairola (020), Rakesh Muchimari ()

Group Code: B21AP01

Index

1. What does it mean?
2. Current Research and methods available?
3. Problems with current methods?
4. Approach taken and problems to address?
5. Possible implementation in real world.

What does it mean by Semantic similarity.

The task of measuring sentence similarity is defined as determining how similar the meanings of two sentences are.

Computing sentence similarity is not a trivial task, due to the variability of natural language expressions. Measuring semantic similarity of sentences is closely related to semantic similarity between words. It makes a relationship between a word and the sentence through their meanings.The intention is to enhance the concepts of semantics over the syntactic measures that are able to categorize the pair of sentences effectively.

Explanation with Example:

Natural Language Processing (**NLP**) field has a term for this, when a word is mentioned we call it a “**surface form**”.

Take for example the word “**president**” by itself this means the head of the country. But depending on context and time it could mean **Trump** or **Obama**.

Query: The economy is more resilient and improving.

1. Microsoft reports strong results as shift to more activities online drives growth in areas from cloud-computing to video games (WSJ) **(Score: 0.5362)**

2. Facebook revenue beats expectations and while ad revenue fell sharply in March there have been recent signs of stability (Bloomberg) **(Score: 0.4632)**

3. Senior White House official confident China will meet obligations under trad deal despite fallout from coronavirus pandemic (WSJ) **(Score: 0.3558)**

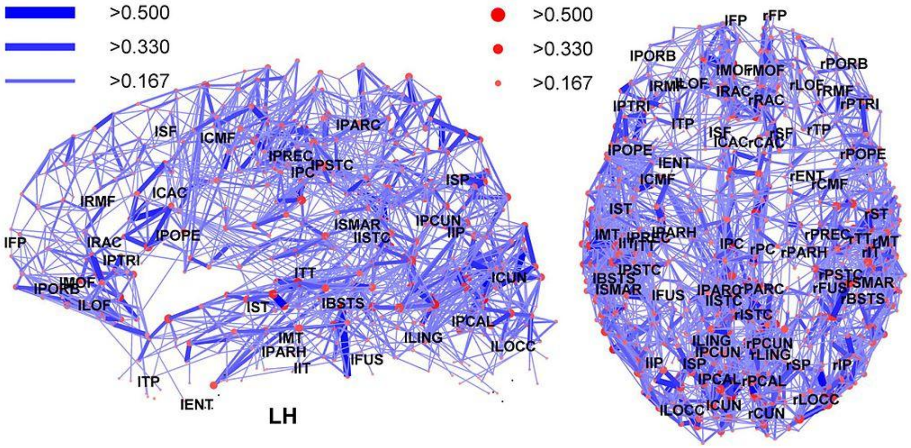
4. Economists from a broad range of ideological backgrounds encouraging Congress to keep spending to combat the coronavirus fallout and don't believe now is time to worry about deficit (Politico) **(Score: 0.3052)**

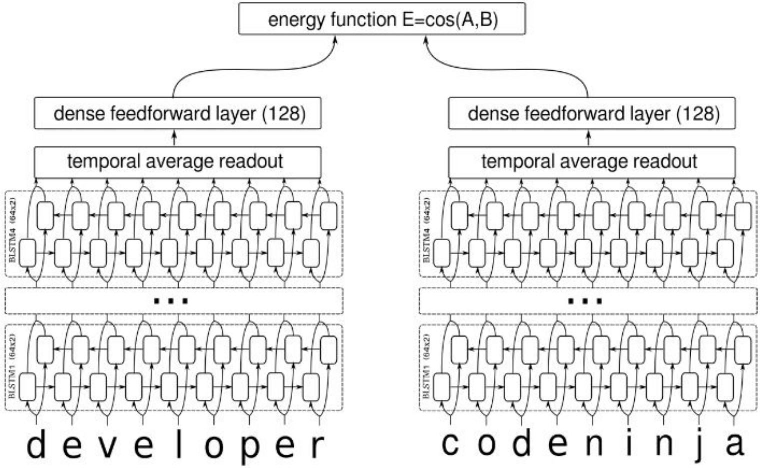
5. White House risks backlash with coronavirus optimism if cases flare up again (The Hill) **(Score: 0.2885)**

model = SentenceTransformer('**bert-base-nli-mean-tokens**')

Current Research available

- Sequence Similarity using Deep LSTM based Siamese Network.
- Sentence comparison with knowledge base (with transformers).
- cosine similarity (Embedding using WordNet).
- path based approach (wu-palmer and shortest path based).
- Feature based approach.





A Closer Look

Approaches and there Limitations.

Sequence Similarity using Deep LSTM based Siamese Network

Link: https://github.com/hrs2203/deep-siamese-text-similarity

Implementation of deep siamese LSTM network to capture phrase/sentence similarity using character embeddings.

Search Spaces Available:

Phrase similarity using **char** level embeddings.

Sentence similarity using **word** level embeddings.

For both the approaches mentioned above it uses a multilayer siamese LSTM network and euclidean distance based contrastive loss to learn input pair similarity.

Limitations:

Computationally expensive

Needs supervision while training in terms of dataset.

Sentence comparison using knowledge base.

Link: http://www.cs.cmu.edu/~mg1/thesis.pdf

Knowledge bases have proven to be incredibly useful for enriching **search results**, **answering factoid questions**, and **training semantic parsers** and **relation extractors**. The way the knowledge base is actually used in these systems, however, is somewhat **shallow**, they are treated most often as simple lookup tables.

Approach we are going for:

Compare how the provided pare off sentences stand with the knowledge base.

Limitations:

Computationally expensive

Needs supervision while training in terms of dataset.

Cosine similarity (Embedding using WordNet)

WordNet is the product of a research project at Princeton University. It is a large lexical database of English. In WordNet nouns, verbs, adverbs and adjectives are organized by a variety of semantic relations into synonym sets(synsets), which represent one concept (Vector).

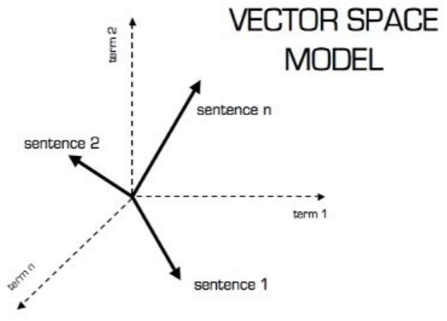
We have a **Vector Space Model** of sentences modeled as vectors with respect to the terms and also have a formula to calculate the similarity between different pair of sentences in this space.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

Limitations:

Knowledge representation at a very abstract level.

Does not guarantees decent output even after training.



Path based approach (wu-palmer and shortest path based)

The Wu & Palmer calculates relatedness by considering the depths of the two **synsets** in the WordNet taxonomies, along with the depth of the **LCS** (Least Common Subsumer).

The formula is **score = 2 * depth (lcs) / (depth (s1) + depth (s2))**.

Limitations:

Knowledge representation at a very abstract level.

Does not guarantees decent output even after training.

Feature based approach

To compute the similarity we follow feature based approach which generates the similarity score in depth of word meaning level and definition level and then comparing the generated results with the previous existing measures for better results. Semantic distance/similarity values of pairs of sentences were calculated using the proposed measure. Therefore, in overall, the proposed measure performs very well and has great potential.

Limitations:

Knowledge representation at a very abstract level.

Does not guarantees decent output even after training.

The Big Question

Which approach to work on and why?

Road Ahead

Our plan for the next 5 months.

What algorithms should we use for our project and how to decide that?

- Reducing our options this early would lead to reduced output quality.
- **Part 1:** (~ 2 months) Implement the above approaches using pytorch and test which one sites out purpose best.
- **Part 2:** (~ 3 months) Based on results from out models, we will look for improvements that we can add on top of those approaches and work from there.

Real World Applications

Search through big corpus of text data if implemented properly has profound effect in real world.

Possible areas where we can leverage this technology are:

1. Historical Scripts.
2. Medical Records.
3. Legal Document query.
4. Search Engine optimization.
5. Fake News detection.

Conclusion

We have to perform lots of experiment with different approaches available and modify them to obtain a generalized and optimized algorithms.

- Thank You