Semantic Similarity Tool

Jéssica Silva 3h ago

- **Goal**: Calculate a semantic similarity score between sentences or paragraphs.
- **Output**: Semantic Similarity score between 0 and 1 at sentence/paragraph-level
- **Setup**: Multilingual

Feature:

- 49420 **[NLP][NLPMetrics] Sentence Semantic Similarity Score** New

Contents

1. Description

The **Semantics Similarity project** aims to provide a high-level interface for calculating a **semantic similarity score** between sentences or paragraphs.

2. Motivation

A **Semantic Similarity score** can be applied to several use cases such as creating or validating datasets.

At DefinedCrowd, we have **Corpora Design projects** that are carried out by crowd via the platform, through the steps of Collection and Validation. We also have **Translation projects** that are done by vendors through the Translation and Validation steps.

Summarizing, this metric can be applied with the aim of:

- Improve the throuput of Workflows through automatic data filtering/inspection
- Increase the quality of deliveries through automatic validation

Below are some examples of possible applications:

Translation projects

- **Quality Estimation**: Assess the quality of a translation pair through the semantic similarity between the source and the target. Remove or improve the worst translations before delivering the dataset to the client (For more details, take a look at our Multilingual Quality Estimation project).

Corpora Design projects

- **Paraphrase Mining**: Task of finding paraphrases (texts with identical/similar meaning) in a large corpus of sentences.
- **Translated Sentence Mining**: Bitext mining describes the process of finding parallel (translated) sentence pairs in monolingual corpora.
- **Similar Texts Retrieval**: Given a document, search a large collection for matching documents.

Domain Clustering projects

- **Clustering**: Identify clusters based on categories or topics. Crowd is only used to disambiguate sentences where the model has low confidence.

Topic Modeling projects

- **Topic Modeling**: Discovering topics in a collection of documents. The Semantic Similarity score can be used to identify topics in a collection of sentences, paragraphs or short documents.

Text Generation/Data Augmentation projects

- **Text Content Validation**: measure the semantic similarity of two texts. If the texts are very similar semantically, this dataset may not have the necessary entropy that the end customer needs to train their models.
- **Text Summarization**: The cosine similarity is calculated across all possible sentence combinations and a ranking algorithm is used to find the most central sentences in the document. These central sentences form a good basis for a summarization of the document.

Thinking about this variety of applications, we decided to work on a project to calculate the **Semantic Similarity score** of **mono** and **multilingual sentences** or **paragraphs**.

3. Project design

The project has two main components:

- The model
- The Tool

3.1 The Model

The model is a **sentence embedding model** and was pretrained in public translation datasets from the OPUS website (http://opus.nlpl.eu/). This model was also fine-tuned on DC translation data (more details about the data).

As embeddings are dense numerical vectors, it is possible to capture the semantic similarity through the cosine distance/similarity between two vectors.

The current model supports 50+ languages: ar, bg, ca, cs, da, de, el, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt-pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw.

The input's limit is ~500 sub-tokens (word-pieces). If the model receives text at the document level, the input will be truncated.

3.2 The Tool

The tool has two features: the **Cosine Similarity Score** and the **Semantic Search**.

3.2.1 Cosine Similarity Score

The **Cosine Similarity feature** measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors (range: [0,1]).

- 0: low similarity
- 1: high similarity

Monolingual examples

Similar sentences:

```
- The new movie is awesome
- The new movie is so great

Score: 0.9821
```

```
- The girl is playing guitar
- The girl is playing a instrument

Score: 0.7536
```

```
- My uncle is the most funny person in the family, and when we all get together, he always make us laugh
- My friend is the most hilarious person in the group, and when we all get together, she always make us laugh

Score: 0.6683
```

```
- Hi Beverly, Thank you for helping Max with his ball passing.
He has been able to pass it better now.
I know that the little lad has got a lot of work to do in offense and defense.
But he is trying his best of him, training everyday.
And with your help, he can become a better player as he grows up.
Thanks again. Ally

- Hi Michele, Thank you for training Henry.
He has been trying really hard to himself on offense and defense.
He also did a lot of cardio to improve his endurance.
His position of him being midfielder is such a big responsibility to him this has inspired him to work hard at getting himself better.
Again, thank you. Mrs Dale, Henry's mom

Score: 0.5597
```

Multilingual examples

Translated sentences (using Google Translator):

en-fr

```
- The new movie is awesome
- Le nouveau film est génial

Score: 0.9772
```

en-it

```
- The girl is playing guitar
- La ragazza suona la chitarra

Score: 0.9567
```

en-ja

```
- My uncle is the most funny person in the family, and when we all get together, he always make us laugh
- 私の叔父は家族の中で最も面白い人です、そして私たち全員が集まるとき、彼はいつも私たちを笑わせます

Score: 0.9112
```

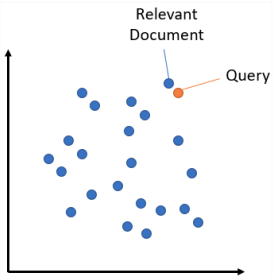
en-ru

```
- Hi Beverly, Thank you for helping Max with his ball passing.
He has been able to pass it better now.
I know that the little lad has got a lot of work to do in offense and defense.
But he is trying his best of him, training everyday.
And with your help, he can become a better player as he grows up.
Thanks again. Ally

- Привет, Беверли! Спасибо за помощь Максу с передачей мяча.
Теперь он смог сдать это лучше.
Я знаю, что у этого маленького мальчика много работы в нападении и защите.
Но он старается изо всех сил, тренируясь каждый день.
И с вашей помощью он может стать лучшим игроком, когда вырастет.
Еще раз спасибо. Союзник

Score: 0.8663
```

3.2.2 Semantic Search

The **Semantic Search feature** is an information retrieval function that focuses on the meaning of the sentences rather than the conventional keyword matching. A information retrieval system is composed by a query and a batch of sentences/paragraphs.

The feature embedded all entries into a vector space. At search time, the query is embedded into the same vector space and the closest embedding from your batch are found. These entries should have a high semantic overlap with the query:



Monolingual examples

Similar Sentences

```
A man is travelling a lot

A man is travelling the world, 0.7379255
A man is travelling to another country, 0.68641114
A woman is travelling to Marrocos, 0.4185376
A man is riding a horse, 0.41770238
A monkey is playing drums, 0.12673867
A cheetah is running behind its prey, 0.104652524
```

```
A cat is running outside

A cat is walking outside, 0.8219033
A animal is prowling outside, 0.59033746
A cat is hunting in the garden, 0.5615754
A tiger is running away from the zoo, 0.5525464
A kitten is sleeping in my bed, 0.4249606
A dog is walking in the garden, 0.3859181
```

Multilingual examples

Similar Sentences (using Google Translator)

en-pt

```
A man is travelling a lot

Um homem está viajando pelo mundo, 0.71998155
Um homem está viajando para outro país, 0.6749014
Uma mulher está viajando para Marrocos, 0.40242425
Um homem está montando um cavalo, 0.34359193
Uma chita está correndo atrás de sua presa, 0.15692018
Um macaco está tocando bateria, 0.15655152
```

en-zh

```
A cat is running outside

一隻貓在外面走, 0.8032492
一隻貓在花園裡打獵, 0.56547815
一隻老虎從動物園逃跑, 0.52713895
動物在外面閒逛, 0.4480929
小貓在我的床上睡覺, 0.40476948
一條狗在花園裡散步, 0.3521912
```

***This same feature can be applied to the Bitext mining task, which describes the process of finding parallel (translated) sentence pairs in monolingual corpora.

4. Usage

4.1 Semantic Similarity Tool as a WEB API

The Semantic Similarity tool is turned into a WEB API by wrapping the script with FastAPI.

To run the server locally see the README.

Once the server is running, the endpoints' documentation will be listed in the swagger UI under `localhost/docs`. Here is a running example https://nlp-semantics.definedcrowd.tech/docs

*** For the time being the Swagger UI does not describe the endpoint schema. It's given below instead.

POST /cosine_similarity

Returns the cosine similarity score between two sentences/paragraphs.

- sentence1: first sentence
- sentence2: second sentence

Request body:

```
{
  "sentence1": "The cat is sleeping outside",
  "sentence2": "The kitten is resting in the garden"
}
```

Response body:

```
{
  "cosine_similarity_score": 0.6746562719345093
}
```

POST /semantic_search

Returns the semantic search for a query (sentence or paragraph) related to a batch of sentences/paragraphs.

- sentence: query
- topn: top N neighbors
- batch: the batch of sentences

Request body:

```
{
  "sentence": "A man is travelling a lot",
  "topn": 3,
  "batch": [
    "A man is travelling to another country",
    "A man is travelling the world",
    "The girl is carrying a baby.",
    "A man is riding a horse",
    "A woman is travelling to Marrocos",
    "Two men pushed carts through the woods",
    "A man is riding a white horse on an enclosed ground",
    "A monkey is playing drums",
    "A cheetah is running behind its prey"
  ]
}
```

Response body:

```
{
  "iterations": [
    {
      "semantic_search": {
        "sentence": "A man is travelling the world",
        "score": "0.7379255"
      }
    },
    {
      "semantic_search": {
        "sentence": "A man is travelling to another country",
        "score": "0.68641114"
      }
    },
    {
      "semantic_search": {
        "sentence": "A woman is travelling to Marrocos",
        "score": "0.4185376"
      }
    }
  ]
}
```

5. Future roadmap

The **Semantic Similarity Tool** can also provide semantic similarity at the word and document level through the adoption of new types of models:

- **word embedding model**
- **document embedding model**

6. Git Repository

Access the project git repository here.

---

31 visits in last 30 days

PB **Pedro Balage** commented Friday

Hi **@Jéssica Silva**,

Here are some suggestion to improve this page. Feel free to apply the suggestions that you do think makes sense. I will erase this comment after closing the story 😃

- Looking for the ASR and VAD documentation, I see that the motivation section plays an important part in guiding the documentation and explaining its contribution to DC problems. I saw that you described the goal and possible applications, but I think you could elaborate more on that. For example, you could describe real use cases and problems we have at DC that could use that (Semantic similarity, domain validation, translation accuracy, etc). You can also mention the benefits of using this tool (improve efficiency in the workflows, increase the delivery quality, inspect data, etc).
- Another topic that is missing in the document is some examples and a space dedicated to explaining multilingual usage. For example, you could add some examples for translation accuracy and/or examples for cross-lingual semantic search. This is a key benefit from this tool and we should promote it more on the page.

Jéssica Silva commented 3h ago

@Pedro Balage thanks for the suggestions. It is much better now.

Add a comment...