Semantic Textual Relatedness (STR)

Rajarshi Dutta | 200762
Shivam Pandey | 200938
Udvas Basak | 201056

Indian Institute of Technology Kanpur

11th November, 2023

Abstract

Why Semantic Textual Relatedness?

- Crucial for deciphering the meaning of text.
- Applied in various NLP tasks, such as Information Retrieval, Question Answering, Text Summarization, Machine Translation, and Paraphrase Detection.

Challenge Overview:

- Focus on automatically detecting the degree of relatedness between pairs of sentences.
- Includes 14 languages, spanning both high-resource and low-resource languages.
- Emphasis on Asian and African languages.

Major Challenge:

- Efficient development of metrics for calculating relatedness scores between sentence pairs.
- Utilizing the linguistic structure of multiple languages to create efficient models.

All files relating to our solution can be found at: https://github.com/Rajarshi1001/CS779AProject.

Problems

Supervised: Submit systems that have been trained **using the labeled training datasets provided**. Allowed to use any **publicly available datasets** (e.g., other relatedness and similarity datasets or datasets in any other languages).

Unsupervised: Submit systems that have been developed without the use of any labeled datasets **pertaining to semantic relatedness or semantic similarity** between units of text more than two words long in any language. **The use of unigram or bigram relatedness datasets (from any language) is permitted**

Cross-lingual: Submit systems that have been developed without the use of any labeled semantic similarity or semantic relatedness datasets in the target language and with **the use of labeled dataset(s) from at least one other language**.

Other papers

What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study (2023) - M. Abdalla et al

Used STR-2022 dataset, annotated for diverse domain-relatedness scores, Fine-tuned BERT-based models, such as BERT-base (mean) and RoBERTa-base (mean), outperform unsupervised models, achieving high Spearman correlations (0.82 and 0.83). Identifies related nouns and objects as influential factors in relatedness.

The Google Similarity Distance (2007) - R.L. Cilibrasi et al

Normalized Google Distance (NGD) utilizes Google search result counts, normalizing co-occurrence considering overall corpus size. Differentiates semantic relatedness from similarity by considering the frequency of words in Google searches. Related terms may appear frequently due to shared context or thematic connections, even if meanings are not identical.

Other papers(contd.)

Making Mono- lingual Sentence Embeddings Multilingual using Knowledge Distillation (2019) - N.Reimers et al

Extends monolingual sentence embeddings to new languages through multilingual knowledge distillation. Using a mean-squared loss function enhances cross-language retrieval performance, outperforming LASER, mUSE, and LaBSE in multilingual STS tasks.

Semantic textual similarity for modern standard and dialectal Arabic using transfer learning (2022) - Al Sulaiman et al

STS in Arabic and dialectical variants (Egyptian, Saudi) using 3 approaches: automatic translation of English STS data, interleaving English STS data with Arabic BERT models, and fine-tuning through knowledge distillation. Dataset translation into Modern Standard, Egyptian, and Saudi Arabic contributes to STS models, yielding a 2% accuracy gain over the SOTA for MSA.

L3Cube-IndicSBERT A simple approach for learning cross-lingual sentence representations using multilingual BERT (2023) - S. Deode et al

Overview: Novel approach to address the scarcity of high-quality language models for low-resource Indian languages. Introduces a range of SBERT models for ten popular Indian languages, trained using a synthetic corpus.

Dataset: English XNLI data translated into 11 Indian languages, for training monolingual SBERT models. Used STS benchmark (STSb) dataset, translated for 10 Indian languages, for evaluating both monolingual and multilingual models on STS.

Novelty: Leveraging multilingual BERT embeddings fine-tuned on a diverse dataset of Indian languages. Introduces a language-specific pooling layer and utilizes AVG instead of CLS pooling.

Results: SOTA performance on cross-lingual sentence similarity tasks for Indian languages, an average accuracy improvement of 8.3% over existing models (from 74.2% to 82.5%).

Labelled Data

The preprocessed data can be found at Kaggle.

Language	Train	Dev
English	5500	250
Moroccan Arabic	1000	118
Amharic	992	95
Hausa	1736	212
Marathi	1200	300
Telugu	1170	130
Spanish	1562	140
Algerian Arabic	924	71
Afrikaans	1761	97

Table: Sizes of Training and Validation Sets for each language

- Only non-semantic variable at the preliminary level is dataset length.
- Correlation coefficients: $-0.13 \leq \rho \leq 0.15$.
- Data is well-distributed and suitable for training.

Plan of Action

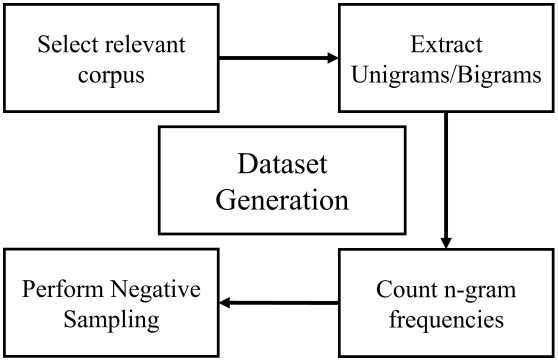
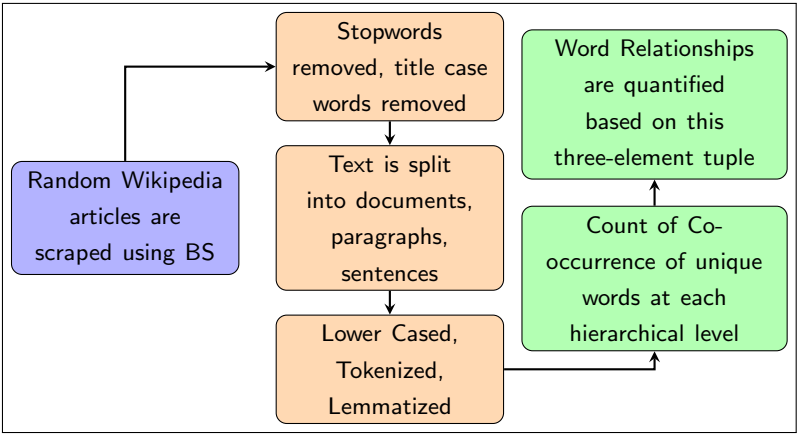
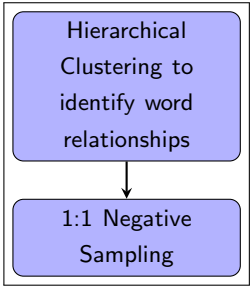


Figure: Data Generation Strategy for Unsupervised Task

1. The algorithm should be able to work on any language(even low-resource).
2. The coverage of the dataset created should be high.
3. The organization in the corpus is set as follows:
   - Corpus
   - Documents
   - Paragraphs
   - Sentences

Implementation for English





Some Observations

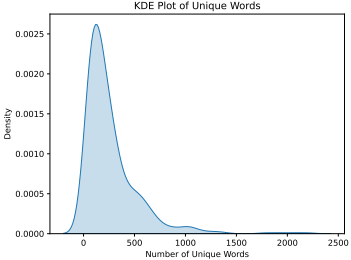


Figure: KDE Plot of number of unique Words in scraped Wikipedia page of each iteration

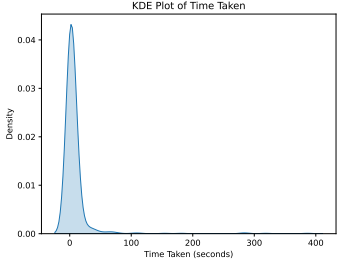


Figure: KDE Plot of time taken for scraping and updating vocabulary in each iteration

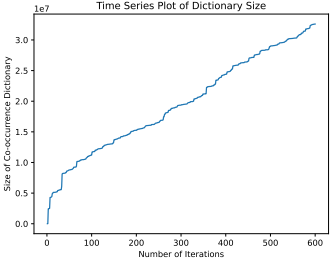


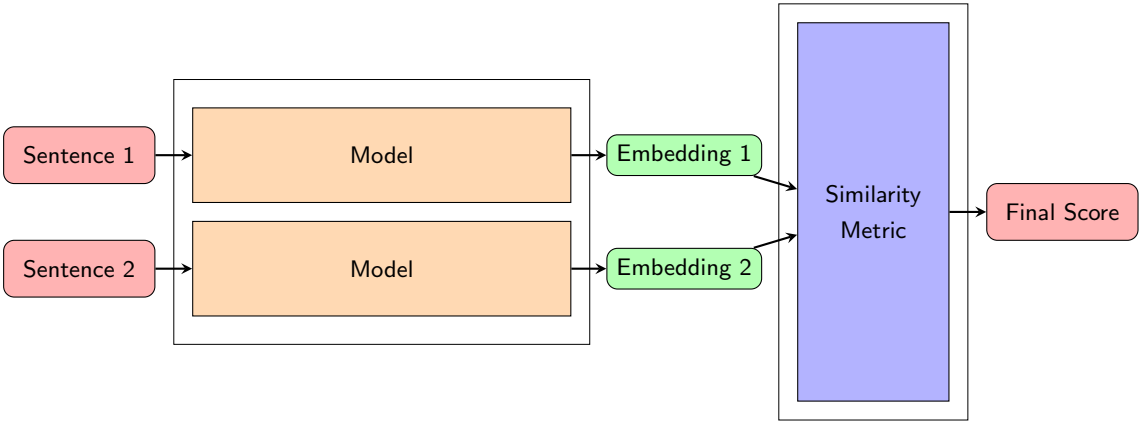
Figure: Evolution of the Bigram Co-occurrence Dictio- nary Size with Iterations

Some Observations(contd.)

1. A corpus was created using 600 randomly sampled Wiki pages.
2. 65.22% of the individual words in the dev set matched with our bigram vocabulary.
3. 53.97% of the bigrams extracted from the dev set matched our bigram vocabulary.
4. 24.21% of the individual words in the dev set were present in our bigram vocabulary, but had never co-occurred in any Wiki article.

1. Since pretrained models were allowed in this track, competitive edge can only be gained through creating unique datasets.
2. Our algorithms can work on any language corpus. Currently trying to implement on Indian Languages.

Overview



Independently tweaking the model, and tweaking the similarity metric.

Experiments on BERT

1. The most naive approach is to fine-tune BERT.
2. One additional step, trying to find exact issues with BERT that can be improved upon
3. One noticeable experiment - scoring with BERT embeddings.
4. Another approach includes fine tuning BERT via vocabulary extension to incorporate tokens of low resource languages

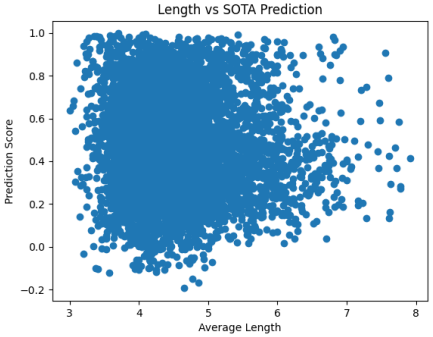


Figure: Average Length vs Prediction Score using sBERT. ρ = -0.0468

mBERT

Hyperparameters	Value
Learning Rate	2e-5
Dropout Rate	0.1
Weight Decay	0.01
Number of linear layers	2
Activation	GELU
Max length	512

Table: Hyperparameters for extended mBERT

Language	Val loss	Val Correlation
Hausa	0.3133	0.0511
Spanish	0.1310	0.5627
Amharic	0.1902	0.4057
English	0.1315	0.7010
Marathi	0.0834	0.0483
Telugu	0.2213	0.3578

Table: Validation stats for some low and medium resource languages on 80-20 dev set

A new similarity metric

1. Multiple metrics can be combined into a single unique metric
2. Focussing on easy to obtain metrics
   - Cosine Similarity
   - Manhattan Dist.
   - Euclidean Dist.
   - Mahalanobis Dist.
   - Dice Coefficient
   - Jaccard Coefficient
3. Vectors element wise powered, and then more metrics are calculated.
4. A total of $42(2 + 4 \times 10)$ metrics used as input to a regression model.
5. $R = 0.8025$ was obtained for English. The model might come handy for other languages.

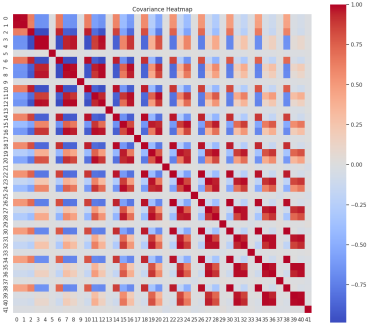


Figure: The higher powers are less correlated, hence more rich

Codalab Correlation Scores on Test Set

Language	Supervised			Unsupervised			Cross-lingual		
	Group-3	Group-4	Group-6	Group-3	Group-4	Group-6	Group-3	Group-4	Group-6
English	0.82	0.82	0.810	-	0.78	0.780	-	-	0.655
Marathi	0.558	0.85	0.750	-	-	0.741	-	-	-
Telugu	0.656	0.83	0.827	-	-	0.800	-	-	-
Hausa	0.401	0.76	0.717	0.273	0.46	0.506	-	-	-
Amharic	0.347	0.83	0.791	0.427	0.508	0.719	-	-	-
Moroccan Arabic	-	0.87	0.742	-	-	0.451	-	-	0.230
Algerian Arabic	-	0.51	0.730	0.398	0.460	0.582	-	-	0.298
Spanish	0.609	-	0.705	0.646	-	0.658	-	-	0.675

Table: Codalab Correlation Scores on Test Set

Future Ideas

Currently experiments in 3 routes:

Low R. → High R. Translator:

- Find(or train) a translator
- Process already underway, trying to translate Hindi-English.
- Possible path for data augmentation, Translate and translate back.
- High potential in Cross-Lingual Problem

Better Models:

- Search and train for better pre-trained models
- HuggingFace already has many models in the unsupervised track.
- Integrating our unsupervised corpus in a clever way.

Better Metrics:

- Mostly a completed path, still looking out for ways to actually implement some non-transformer relatedness metric(NGD)
- Metrics thus developed might help in other tracks, potential for cross-lingual
- Autoencoding and metric calculation