Project#3: Semantic Textual Relatedness

Rajarshi Dutta[1], **Udvas Basak**[2], **Shivam Pandey**[3]
[1]200762, [2]201056, [3]200938
[1]MSE, [2]AE, [3]ME
rajarshi20@iitk.ac.in, udvasb20@iitk.ac.in, shivamp20@iitk.ac.in

Semantic Textual Relatedness is important for deciphering meaning and its applications lie on various NLP tasks like **Information Retrieval**, **Question Answering**, **Text Summarization**, **Machine Translation**, **Paraphrase Detection** etc. The challenge is mainly focussed on automatically detecting the degree of relatedness between pairs of sentences, for 14 languages, which include high-resource as well as low-resource languages, specially Asian and African languages. The major challenge lies in the efficient development of a metric to facilitate the calculation of the **relatedness** score between the sentence pairs, and harnessing the structure of multiple languages to create an efficient model. The data is trained on pairs of sentences with manually determined relatedness scores between 0(completely unrelated), and 1(maximally related).

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

1 Introduction

Semantic relatedness measures the proximity in meaning between units of language, such as words, sentences, phrases, or n-grams (Mohammad, 2008). Semantically similar sentences exhibit paraphrasal or entailment relations, while relatedness encompasses broader commonalities, including topic alignment, viewpoint, temporal context, elaboration, and more (e.g., two sentences about the same event). Most NLP work focuses on Semantic Textual Similarity in English, emphasizing the importance of the scoring metric. Relatedness algorithms like Normalized Google Distance (Lopes and Moura, 2019), which do not rely on transformers, show promise in calculating relatedness scores.
This paper documents our efforts in addressing this problem, particularly in combining relatedness scores intelligently to achieve high accuracy.

2 Problem Definition

The task presented in SemEval 2024 has three tracks. The tasks are presented verbatim as follows:

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

Unsupervised: Participants are to 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: Participants are to 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**.

3 Related Work

Key points:

1. (Reimers and Gurevych, 2019) Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation - Nils Reimers, Iryna Gurevych Ubiquitous
   This study introduces a novel approach to extend monolingual sentence embedding models to new languages. It uses multilingual knowledge distillation, where a student model, $\hat{M}$, learns from a

teacher model, M, on source language sentences (s_i) and their translations (t_i) by minimizing the mean-squared loss function.

$\text{Loss}=\frac{1}{|B|}\sum_j \big(M(s_j)-\hat{M}(s_j)\big)^2+\big(M(s_j)-\hat{M}(t_j)\big)^2$

This method significantly boosts retrieval performance across languages, especially for low-resource languages, surpassing LASER, mUSE, and LaBSE in multilingual semantic textual similarity (STS) tasks. It successfully aligns vector spaces across languages, addressing limitations in LASER and LaBSE, but also reveals a "curse of multilinguality," where adding more languages can harm model performance.

Future research may explore additional languages and domains to assess generalizability. Mitigating language bias in multilingual models is another promising direction.

2. The Google Similarity Distance - Rudi L. Cilibrasi, Paul M.B. Vitanyi

(Lopes and Moura, 2019) Cilibrasi and Vitanyi introduced the concept of Normalized Google Distance (NGD) as a novel metric to measure semantic similarity between words or concepts. The NGD is calculated using Google search result counts, enabling a data-driven approach to quantify the relatedness of terms. Their formula for NGD is given by:

$\text{NGD}(w_1,w_2)=\frac{\max(\log(N(w_1)),\log(N(w_2)))-\log(N(w_1,w_2))}{\log(N)-\min(\log(N(w_1)),\log(N(w_2)))}$

In this formula, N represents the total number of web pages indexed by Google, N($w1$) and N($w2$) are the counts of pages containing words $w1$ and $w2$, and N($w1, w2$) is the count of pages containing both w1 and w2. The NGD formula normalizes these counts by considering the overall corpus size and the co-occurrence of terms in web pages.

Their research confirmed NGD's effectiveness in word sense disambiguation, concept classification, and natural language translation, with consistently high accuracy. Comparing NGD-based semantic categories with WordNet showed strong agreement, highlighting its potential for advancing computational semantics and natural language understanding through web-scale data.

Future research should refine NGD, address biases, extend to sentence analysis, develop scalable algorithms, explore multimodal semantics, and adapt NGD for industry-specific applications.

4 Corpus/Data Description

4.1 Supervised and Cross-Lingual Tracks

Out of the 14 languages, namely, Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu, data has been released only for English, Marathi, Telugu, Hausa, Moroccan Arabic and Amharic.

The sizes of the datasets are noted in Table (1). At

	Train	Dev
English	5500	250
Moroccan Arabic	1000	118
Amharic	992	95
Hausa	1736	212
Marathi	1200	300
Telugu	1170	130

Table 1: Sizes of Training and Validation Sets for each language

the preliminary level, the only non-semantic variable in these datasets is dataset length. To assess semantic relatedness, it's crucial to mitigate these biases. Plots indicate no discernible correlation between sentence lengths and scores, suggesting well-distributed data suitable for training. Correlation coefficients fall in the range $-0.13 < \rho < 0.15$.

Since the English dataset is large, we can do some more preliminary analysis on that data. Particularly, we are interested to see if there is any relation of the current SOTA models [sBERT(Reimers and Gurevych, 2019)] with the length of the sentences. The plot for that can be seen in Fig. (1). Visibly, there is no correlation between them.

4.2 Unsupervised Track

1. For this track, we can use only unigram and bigram datasets. The search for such datasets is ongoing. The availability of such datasets is low, since, most tasks have focussed on semantic textual similarity.
2. A plan for generating a dataset is also ongoing, which will emphasize on semantic relatedness. Fig. (2) shows the tentative plan for developing the required unigram and bigram dataset.

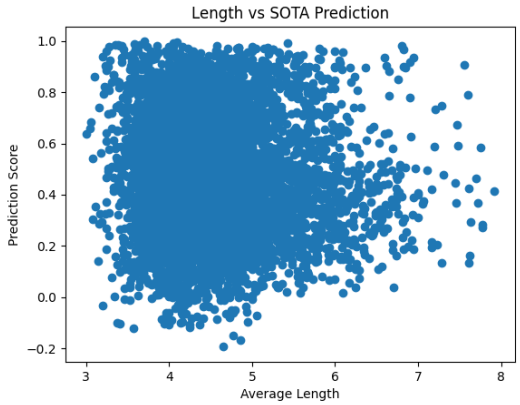

Figure 1: Average Length vs Prediction Score using sBERT. $\rho = -0.0468$

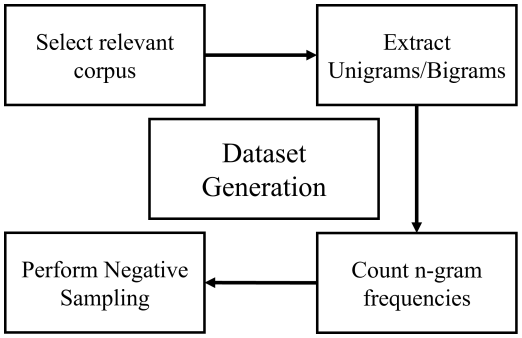

Figure 2: Data Generation Strategy for Unsupervised Task

5 Future Directions

1. Extending NGD for relatedness calculation among sentences

   The algorithm calculates the Normalized Google Distance (NGD) for two input sentences, text1 and text2. It starts by tokenizing and removing stop words from both sentences, followed by part-of-speech tagging. Then, it calculates NGD values for pairs of words with the same part of speech in both sentences. The NGD scores are normalized and averaged to compute the overall NGD score, representing the semantic similarity between the two sentences.

2. Our approach includes utilizing cardinality based metrics like **Dice coefficient** and **Jaccard Similarity**, and similarity and distance metrics like **Cosine similarity**, **Euclidean distance** and **Mahalanobis distance** on preprocessed tokens generated from the sentence pairs as well as devising a **transformer** based architecture for obtaining relatedness scores from the sentence embeddings produced.

3. sBERT based implementaion

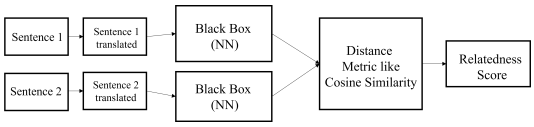

Figure 3: Sentence BERT based approach

4. Siamese Architecture based implementation

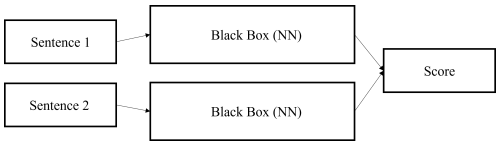

Figure 4: Siamese architecture

5. Initially each of the sentences are passed through a translation system to handle the multilingual aspect of the problem statement. The translated sentences are passed through as **sentenceBERT** to generate sentence embeddings. As for the next part, one approach includes efficiently implementing a **feed-forward neural network** which outputs the relatedness scores in the range of 0 and 1 from the two sentence embeddings. The other approach relies on implementation of an end-to-end **Siamese architecture** trained on the directly outputing the relatedness score without using metrics.

6. Searching for a good relatedness metric as of now or improving the performance of the system by tuning the parameters of the feedforward model to immitate a good relatedness metric system since **relatedness** metric is harder ot design as compared to a **semantic similarity** metric

7. The **supervised** task relies on:

   - Transformer based implementation (described above) for being trained on the labelled data provided in 14 different languages.

   The **unsupervised** task relies on:

- Extraction of bigrams and training a siamese architecture on those bigrams along with some negative samples created using **negative sampling** to generate bigram embeddings
- These trained bigram embeddings are passed through an evaluation metric which are further clustered using methods like **DBSCAN**, **Hierarchial clustering** to calculate the relatedness score based on a threshold value.

The **Cross lingual** task is still under development.

8. Our plan of action currently includes tweaking the architecture parameters and choosing some good models to improve the relatedness score above a baseline which is **0.83**
9. The expected timeline for the future tasks in the project is mentioned below:

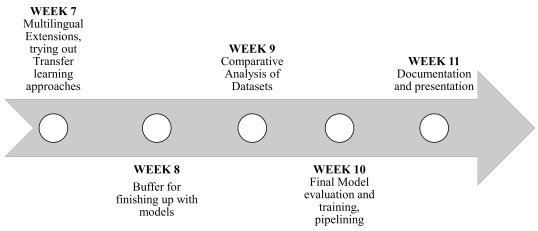

Figure 5: Expected Timeline

6 Individual Contribution

Rajarshi Dutta - Implementation of the BERT based approach to calculate the relatedness score using a feedforward neural network
Shivam Pandey - Searching and navigating through several research papers for good relatedness metrics and models
Udvas Basak - Trying out and improving on traditional lexical overlap and distance based metrics on the datasets to achieve a baseline score

7 Conclusion

In the evolving landscape of **Natural Language Processing**, the distinction and relationship between **semantic similarity** and **semantic relatedness** are paramount. This report underscored the nuanced differences between the two, highlighting the broader spectrum of relatedness in comparison to similarity. With the majority of previous research primarily concentrating on Semantic Textual Similarity and English language, the **SemEval 2024 challenge** presents a timely opportunity to delve deeper into the relatively less charted waters of semantic relatedness across multiple dimensions: **supervised**, **unsupervised**, and **cross-lingual**. Our objectives, as elucidated, steer towards developing comprehensive methods for computing semantic relatedness scores, harnessing both established and novel datasets, and adapting these methodologies to a cross-lingual setting thereby enhancing the richness of machine comprehension in diverse linguistic landscapes.

8 Presentation Feedback (Max 2 columns)

1. Questions include how do we plan to use the transformer and non-transformer based approaches and how do we exactly tackle the unsupervised portion of the problem statement?

References

Carla Teixeira Lopes and Diogo Moura. 2019. Normalized google distance in the identification and characterization of health queries. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.