HausaNLP at SemEval-2024 Task 1: Textual Relatedness Analysis for Semantic Representation of Sentences

Saheed Abdullahi Salahudeen[1,2], Falalu Ibrahim Lawan[2], Aliyu Yusuf[3], Amina Abubakar Imam[4], Lukman Aliyu, Nur Bala Rabiu[5], Mahmoud Said Ahmad, Idi Mohammed[6], Aliyu Rabiu Shuaibu[7], Alamin Musa, Auwal Shehu Ali[8], Zedong Nie[1]

[1]Shenzhen Institute of Advanced Technology, CAS, [2]Kaduna State University,
[3]Universiti Teknologi PETRONAS, [4]University of Abuja, [5]Khalifa Isyaka Rabiu University,
[6]AUST, Abuja [7]Nile University, [8]Bayero University Kano, [∀]HausaNLP.

Contact: zd.nie@siat.ac.cn

Abstract

Semantic Text Relatedness (STR), a measure of meaning similarity between text elements, has become a key focus in the field of Natural Language Processing (NLP). We describe SemEval-2024 task 1 on Semantic Textual Relatedness featuring three tracks: supervised learning, unsupervised learning and cross-lingual learning across African and Asian languages including Afrikaans, Algerian Arabic, Amharic, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu. Our goal is to analyse the semantic representation of sentences textual relatedness trained on mBert, all-MiniLM-L6-v2 and Bert-Based-uncased. The effectiveness of these models is evaluated using the Spearman Correlation metric, which assesses the strength of the relationship between paired data. The finding reveals the viability of transformer models in multilingual STR tasks.

1 Introduction

The rapid increase in digital information has presented a critical challenge for researchers. The web hosts around 50 million pages of text, which is beyond the capacity of human interpretation alone. To interpret this extensive text data effectively, it is essential to comprehend the meanings of various words (Jain et al., 2020). Semantic Text Relatedness (STR) is a semantic analysis of the relationship between two pieces of text based on their meanings. STR of two language units has long been considered fundamental to understanding meaning (Miller and Charles, 1991; Lastra-Díaz and García-Serrano, 2015), It's a metric used to measure the similarity in meaning between two terms or documents. It is a subset of computational linguistics and one of the fundamental concepts of Natural Language Processing (NLP). STR can be measured using datasets designed by experts, which are made up of word pairs that are known to be related. It can be used in identifying a paraphrase or duplicate, as well as search engines to give users relevant and personalized results.

When two sentences have a paraphrase or entailment relation, they are considered to be semantically similar and When evaluating the semantic relatedness between them, humans typically focus on identifying shared meanings. In the case of the sentence pairs below, most English speakers would agree that the sentences in the first pair are more closely related in meaning than those in the second pair, whether they are from the same topic, express the same view or originate from the same time period etc.(Abdalla et al., 2023).

Pair 1: a. *There was a lemon tree next to the house.*
b. *The boy enjoyed reading under the lemon tree.*

Pair 2: a. *There was a lemon tree next to the house.*
b. *The boy was an excellent football player.*

Previous NLP research has mainly dealt with semantic relatedness primarily in English language. However, in this task, we address a variety of languages, including Afrikaans, Algerian Arabic, Amharic, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu. The task featured the following tracks: Track A which is a supervised learning, track B is an unsupervised learning and Track C is a cross-lingual learning.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 188–192
June 20-21, 2024 ©2024 Association for Computational Linguistics

2 Related Works

Sentences are considered semantically related when they share commonalities in meaning, such as paraphrasal or entailment relations. A study by (Abdalla et al., 2023) developed a Semantic Textual Relatedness dataset (STR-2022) to manually annotate English sentence pairs and explore the factors that contribute to the semantic relatedness of sentences. The dataset has been used to study the degree of semantic relatedness and the reliability of human intuition in determining the relatedness of sentence pairs while (Hasan et al., 2020) assessed the methods for semantic relatedness between words based on knowledge sources. These methods exploit features from both structural and statistical approaches, emphasizing on semantic representation, measures of semantic similarity, and knowledge-based text mining.
(Lastra-Díaz and García-Serrano, 2015) proposed Explicit Semantic Analysis (ESA), a recently introduced approach that signifies the meaning of texts by computing the semantic relevance of natural language texts. This approach assumes the need for substantial amounts of common sense and domain-specific knowledge, utilizing machine learning techniques to explicitly depict the meaning of any text. This is achieved by creating a weighted vector based on concepts from Wikipedia. ESA undergoes continuous development, ensuring a consistent expansion of its breadth and depth over time.

3 Task Description

STR Shared Task 1 (Ousidhoum et al., 2024b) consists of predicting the semantic relatedness of sentence pairs. Sentence pairs will be rank based on their closeness in meaning in 14 different languages. All sentence pairs will have manually determined relatedness scores between 0 (completely unrelated) and 1 (maximally related). Participants are provided with a gold label scores with a comparative annotation approach that led to a high reliability of the final relatedness rankings. The shared task consists of three tracks: supervised learning, unsupervised learning and cross-lingual learning. In this paper, we concentrate on all the three tracks.

3.1 Track A: Supervised

This track relies on labelled input and output training data. We used the labeled training datasets for 9 languages provided for the shared task which include: Algerian Arabic, Amharic, English, Hausa, Kinyarwanda, Marathi,Moroccan Arabic, Spanish and Telgu.

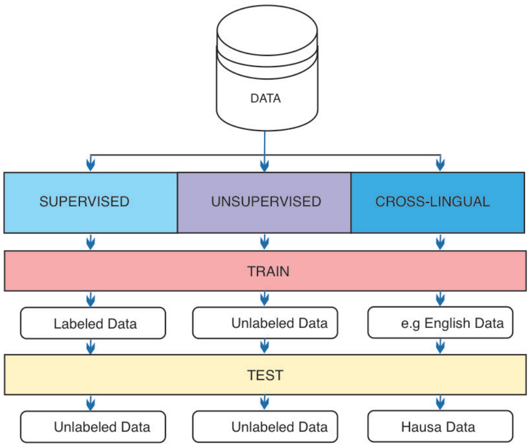


Figure 1: **Task Overview**

3.2 Track B: Unsupervised

Unsupervised learning analyzes and cluster unlabeled datasets, it is typically used when the goal is to identify patterns and relationships in data. We make this analysis using 12 languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi,Indonesian, Kinyarwanda, Modern Standard Arabic, Moroccan Arabic, Punjabi and Spanish.

3.3 Track C: Cross-lingual

Cross-lingual learning involves transferring models from one language to another, typically to improve performance. For this track we make use of 12 languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi,Indonesian, Kinyarwanda, Modern Standard Arabic, Moroccan Arabic, Punjabi and Spanish.

4 Experiment and Evaluation

This section describes the system overview which comprises the dataset description, model description and evaluation metric.

4.1 Dataset Description

The dataset consists of an instance of a sentence pair of both the training, development and test sets. Each instance is annotated with a gold label score that represents the degree of semantic text relatedness between two sentences (Ousidhoum et al., 2024a). The gold label scores are determine by manual annotation and range from 0 (not related at all) to 1 (very related at all). A comparative annotation approach is used to avoid biases of the traditional rating scales and can result to a high reliability of final relatedness rankings. The dataset used in this shared task are from the following languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Morrocan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu.

4.2 Models Description

We experiment with multiple pre-trained models before deciding to go with the selected models based on the tracks. However, due to time constraint and resources, we reported for the competitive models across various languages based on the task specification.

4.2.1 mBERT

We used mBERT in a supervised aapproach, mBert is a multilingual derivative of BERT and trained on a diverse set of 104 languages. The pre-training process for mBERT involves masked language modeling (MLM) and the next-sentence prediction task (Libovickỳ et al., 2019). To tailor the model for our specific task, we fine-tune the mBERT-base-cased model, which boasts 172 million parameters. A 70-30 train-test split is executed with a learning rate of 1e-5 on Adam optimizer.

4.2.2 all-MiniLM-L6-v2

The all-MiniLM-L6-v2 model was used in an unsupervised approach in this task, it is a lightweight transformer-based model for semantic similarity comparison with optimized model size and faster inference (Wang et al., 2020). It has 66 Million Parameters compressed in a Student-Mimicking-Teacher network relationship. Using self attention distribution, we utilized the Teacher's last layer to guide the training of the student distillation in an unsupervised manner and generated effective and flexible results for the 12 languages used.

4.2.3 BERT-BASED-UNCASED

The Bert-Based-Uncased model was used in a cross-lingual approach in this task. It is a pre-trained autoencoding language model trained on vast English Wikipedia and BookCorpus with a sequence length of 512. The model is based on the architecture presented in (Devlin et al., 2018). As the track description, some of the languages were initially trained on different language before applying task on new language. Bert-Based-Uncased use WordPiece tokenizer, it has 110 parameters 12-layer, 768-hidden, 12- attention heads.

Task	Model	Language	Sp. Corr.
Track A: Supervised	mBERT	Algerian Arabic	0.388
		Amharic	0.269
		English	0.762
		Hausa	0.580
		Kinyarwanda	0.527
		Marathi	0.811
		Moroccan Arabic	0.696
		Spanish	0.696
		Telugu	0.791
Track B: Unsupervised	all-MiniLM-L6-v2	Afrikaans	0.468
		Algerian Arabic	0.398
		Amharic	0.098
		English	0.825
		Hausa	0.273
		Hindi	0.465
		Indonesian	0.384
		Kinyarwanda	0.131
		Modern Standard Arabic	0.200
		Moroccan Arabic	0.496
		Punjabi	0.011
		Spanish	0.603
Track C: Cross-Lingual	BERT-BASED-UNCASED	Afrikaans	0.710
		Algerian Arabic	0.780
		Amharic	0.660
		English	0.780
		Hausa	0.630
		Hindi	0.740
		Indonesian	0.790
		Kinyarwanda	0.750
		Modern Standard Arabic	0.660
		Moroccan Arabic	0.670
		Punjabi	0.730
		Spanish	0.810

Table 1: Results of various tasks.

4.3 Spearman Correlation

The Spearman Correlation is a non parametric and normality for monotonic relationship between variables (Ali Abd Al-Hameed, 2022). It measures the strength of relationship between paired data. It is similar to Pearson's Product Moment Correlation Coefficient (De Winter et al., 2016), or Pearson's r. It indicates magnitude and direction of the association between two variables that are on interval or ratio scale. For this task, we used Spearman Correlation to measure the similarity between two sentences.

5 Results and Discussion

This section presents the results of the Shared task Tracks. Table 1 displays the Spearman correlation scores for the evaluation of 14 low-resource languages for semantic relatedness. The SemEval-2024 task on STR provided an opportunity to explore the effectiveness of transformer models. The models capture semantic relatedness across multiple languages. In This section, the analyses and interpretations of the results obtained from the given tasks are Task A (supervised learning), Task B (unsupervised learning) and Task C (cross-lingual).

In supervised learning track A, the multilingual BERT (mBERT) model was used. The model demonstrated different levels of performance across the languages. Notably, mBERT exhibited strong correlation scores in languages such as English with 0.76, Marathi with 0.81, and Telugu with 0.79 correlation. This indicates the model's ability to generalize well across linguistic contexts in semantic relatedness tasks. These findings suggest that mBERT can effectively capture semantic relatedness, even in low-resource languages, highlighting its robustness and cross-lingual generalization capabilities. However, challenges were observed in languages with complex morphological structures, underscoring the need for further research to address such linguistic nuances.

Conversely, the unsupervised learning track B featured the All-MiniLM-L6-v2 model, which achieved promising results in certain languages, particularly English with 0.82, Spanish with 0.60, and Moroccan Arabic with 0.5 Spearman correlation value. Despite its effectiveness, the model faced difficulties in languages such as Punjabi and Amharic, where semantic relatedness was harder to capture without labelled data. These challenges emphasize the importance of developing techniques to improve unsupervised learning models' performance, especially in low-resource language settings.

Similarly, track C (cross-lingual) which were entirely trained with BERT-BASED-UNCASED performed promisingly despite training and predicting on different language pairs. The Spearman correlation for Spanish achieved 0.81 and was trained on English, while Hausa achieved the lowest with 0.63 despite being trained on Kinyarwanda training dataset. This performance especially in Semantic Textual Relationship shows that cross-lingual hold a prospective future for generalization of NLP tasks.

However, the findings highlight the effectiveness of transformer models, in capturing semantic relatedness across diverse languages. The choice of evaluation metrics, such as Spearman correlation, proved instrumental in assessing the models' performance and understanding their ability to capture the ordinal relationship between predicted and true semantic relatedness scores. Furthermore, the results contribute valuable insights into advancing the understanding and application of semantic textual relatedness in multilingual NLP tasks, paving the way for future research in this domain.

6 Conclusion and Future Works

The study on Semantic Text Relatedness (STR) across multiple languages has demonstrated the effectiveness of transformer models in capturing semantic relatedness. The multilingual BERT (mBERT) model showed strong correlation scores in languages such as English, Marathi, and Telugu, indicating its ability to generalize well across linguistic contexts. The All-MiniLM-L6-v2 model achieved promising results in English, Spanish, and Moroccan Arabic, while facing challenges in languages like Punjabi and Amharic. The cross-lingual track, using BERT-BASED-UNCASED, also performed well, especially in Spanish, trained on English data. These findings underscore the potential of transformer models in NLP tasks and the importance of appropriate evaluation metrics like Spearman Correlation. The study contributes valuable insights into advancing semantic textual relatedness in multilingual NLP, highlighting areas for future research and development.

Future work should focus on exploring advanced transformer Large Language Models (LLMs) like GPT-3 and T5 to improve performance across diverse languages, including low-resource ones. Expanding language coverage, incorporating contextual and cultural information, and fine-tuning with language-specific data will enhance model accuracy. Cross-lingual transfer learning techniques can be investigated to adapt high-resource language models to low-resource settings. Hybrid approaches combining different learning methods may offer improved results, while new evaluation metrics could better capture semantic nuances. Additionally, exploring multimodal STR and applying research findings to real-world applications will increase the practical impact of STR systems.

Acknowledgements

We would like to acknowledge HausaNLP Management for their tireless support.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguisticst doi =.

Khawla Ali Abd Al-Hameed. 2022. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 13(1):3249–3255.

Joost CF De Winter, Samuel D Gosling, and Jeff Potter. 2016. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ali Muttaleb Hasan, Noorhuzaimi Mohd Noor, Taha H Rassem, and Ahmed Muttaleb Hasan. 2020. Knowledge-based semantic relatedness measure using semantic features. *International Journal*, 9(2).

Shivani Jain, KR Seeja, and Rajni Jindal. 2020. A new methodology for computing semantic relatedness: modified latent semantic analysis by fuzzy formal concept analysis. *Procedia Computer Science*, 167:1102–1109.

Juan J Lastra-Díaz and Ana García-Serrano. 2015. A novel family of ic-based similarity measures with a detailed experimental survey on wordnet. *Engineering Applications of Artificial Intelligence*, 46:140–153.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

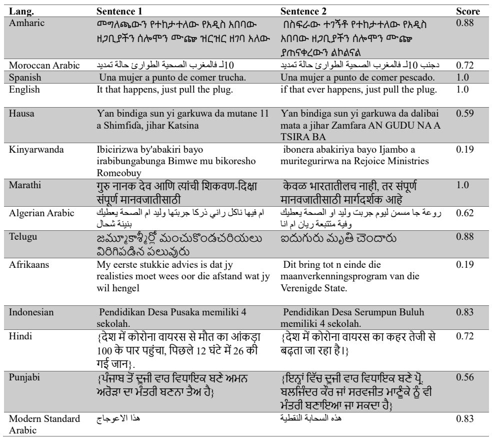
George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

A Appendix



Lang.	Sentence 1	Sentence 2	Score
Amharic	መግለጫውን የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ዝርዝር ዘገባ አለው	በስፍራው ተገኝቶ የተከታተለው የአዲስ አበባው ዘጋቢያችን ሰሎሞን ሙጬ ያጠናቀረውን ልኮልናል	0.88
Moroccan Arabic	تمديد حالة الطوارئ الصحية فالمغرب لـ10	تمديد حالة الطوارئ الصحية فالمغرب لـ10 دجنب	0.72
Spanish	Una mujer a punto de comer trucha.	Una mujer a punto de comer pescado.	1.0
English	It that happens, just pull the plug.	if that ever happens, just pull the plug.	1.0
Hausa	Yan bindiga sun yi garkuwa da mutane 11 a Shimfiɗa, jihar Katsina	Yan bindiga sun yi garkuwa da dalibai mata a jihar Zamfara AN GUDU NA A TSIRA BA	0.59
Kinyarwanda	Ibicirizwa by'abakiri bayo irabibungabunga Bimwe mu bikoresho Romeobuy	ibonera abakiriya bayo Ijambo a muritegurirwa na Rejoice Ministries	0.19
Marathi	गुरु नानक देव आणि त्यांची शिकवण-दिक्षा संपूर्ण मानवजातीसाठी	केवळ भारतातीलच नाही, तर संपूर्ण मानवजातीसाठी मार्गदर्शक आहे	1.0
Algerian Arabic	ام فيها ناكل راني ذركا جربتها وليد ام الصحة يعطيك بنينة شحال	روعة جا مسمن ليوم جربت وليد او الصحة يعطيك وفية متتبعة ريان ام انا	0.62
Telugu	జమ్మూకాశ్మీర్లో మంచుకొండచరియలు విరిగిపడిన పలువురు	ఐదుగురు మృతి చెందారు	0.88
Afrikaans	My eerste stukkie advies is dat jy realisties moet wees oor die afstand wat jy wil hengel	Dit bring tot n einde die maanverkenningsprogram van die Verenigde State.	0.19
Indonesian	Pendidikan Desa Pusaka memiliki 4 sekolah.	Pendidikan Desa Serumpun Buluh memiliki 4 sekolah.	0.83
Hindi	{देश में कोरोना वायरस से मौत का आंकड़ा 100 के पार पहुंचा, पिछले 12 घंटे में 26 की गई जान}.	{देश में कोरोना वायरस का कहर तेजी से बढ़ता जा रहा है।}	0.72
Punjabi	{ਪੰਜਾਬ ਤੋਂ ਦੂਜੀ ਵਾਰ ਵਿਧਾਇਕ ਬਣੇ ਅਮਨ ਅਰੋੜਾ ਦਾ ਮੰਤਰੀ ਬਣਨਾ ਤੈਅ ਹੈ}	{ਇਨ੍ਹਾਂ ਵਿੱਚ ਦੂਜੀ ਵਾਰ ਵਿਧਾਇਕ ਬਣੇ ਪ੍ਰੋ. ਬਲਜਿੰਦਰ ਕੌਰ ਜਾਂ ਸਰਵਜੀਤ ਮਾਣੂੰਕੇ ਨੂੰ ਵੀ ਮੰਤਰੀ ਬਣਾਇਆ ਜਾ ਸਕਦਾ ਹੈ}	0.56
Modern Standard Arabic	هذا الاعوجاج	هذه السحابة النقطية	0.83

Figure 2: Example Sentences