NLU-STR at SemEval-2024 Task 1: Generative-based Augmentation and Encoder-based Scoring for Semantic Textual Relatedness

Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia
Birzeit University, Palestine
{smalaysha, mjarrar, mkhalilia}@birzeit.edu

Abstract

Semantic textual relatedness is a broader concept of semantic similarity. It measures the extent to which two chunks of text convey similar meaning or topics, or share related concepts or contexts. This notion of relatedness can be applied in various applications, such as document clustering and summarizing. SemRel-2024, a shared task in SemEval-2024, aims at reducing the gap in the semantic relatedness task by providing datasets for fourteen languages and dialects including Arabic. This paper reports on our participation in Track A (Algerian and Moroccan dialects) and Track B (Modern Standard Arabic). A BERT-based model is augmented and fine-tuned for regression scoring in supervised track (A), while BERT-based cosine similarity is employed for unsupervised track (B). Our system ranked 1st in SemRel-2024 for MSA with a Spearman correlation score of 0.49. We ranked 5th for Moroccan and 12th for Algerian with scores of 0.83 and 0.53, respectively.

1 Introduction

The literature commonly examines semantic similarity, where the focus is on whether two linguistic units (words, phrases, sentences, etc.) share similar meanings (Bentivogli et al., 2016). However, semantic textual relatedness (STR) is less explored due to its complexity and the scarcity of datasets (Abdalla et al., 2023; Darwish et al., 2021). While the former task checks for the presence of similar meaning or paraphrase, STR takes a more comprehensive approach, evaluating relatedness across multiple dimensions, spanning topical similarity, conceptual overlap, contextual coherence, pragmatic connection, themes, scopes, ideas, stylistic conditions, ontological relations, entailment, temporal relation, as well as semantic similarity itself (Miller and Charles, 1991; Halliday and Hasan, 2014; Jarrar, 2021, 2011). For example, consider the two sentences (*The Earth orbits the sun at a speed of ~110,000 km/h.*) and (*Earth rotates at ~1670 km/h around its axis.*). They hold semantic relatedness through the shared topic of Earth's speeds. In contrast, both sentences are not semantically similar as they possess distinct meanings. This illustrates the broader range of STR as described by Abdalla et al. (2023), which ranges from highly relevant sentences, expressing the same idea with different wording, to entirely unrelated sentences, discussing unrelated topics.

Semantic relatedness has proven to be useful in evaluating sentence representations generated by language models (Asaadi et al., 2019), in addition to question answering (Tsatsaronis et al., 2014), machine translation (Mi and Xie, 2024), plagiarism detection (Sabir et al., 2019), word-sense disambiguation (Al-Hajj and Jarrar, 2021a; Malaysha et al., 2023), among others. Exploring the relatedness and similar tasks in languages other than English is hindered by the lack of data (Jarrar et al., 2023b; Al-Hajj and Jarrar, 2021b). The SemRel-2024 shared task (Ousidhoum et al., 2024a) provided datasets in fourteen languages and offered three tracks. In the supervised track (A), training and testing are performed on the same language. In the unsupervised track (B), the use of labeled data for training is prohibited; and in the cross-lingual track (C), testing is conducted on a different language than the one used for training.

This paper presents our contribution to track A and track B. In track A, we fine-tuned BERT models using the Algerian and Moroccan sentence pairs to produce similarity scores. To enrich the data, we augmented the SemRel-2024 dataset (Ousidhoum et al., 2024a) by generating additional sentence pairs from Google Gemini [1], a generative model, using a predefined prompt template. These generated pairs imitated the style and meaning of the existing pairs, and we assigned them

[1] https://gemini.google.com/

scores corresponding to the originals. We used the same datasets provided by the Shared Task in addition to a ~760 augmented Moroccan pairs to fine-tune BERT models, AraBERTv2 (Antoun et al., 2020) and ArBERTv2 (Abdul-Mageed et al., 2021), which resulted in a performance enhancement of 0.05 points. In track B, as training on labeled data is not allowed, we used cosine similarity using average pooling embedding (Zhao et al., 2022) on top of each model. Our approaches achieved Spearman scores (Tsatsaronis et al., 2014) of 0.49 for MSA (ranked first), 0.83 for Moroccan (ranked fifth), and 0.53 for Algerian (ranked twelfth).

2 Related Work

Semantic textual relatedness (STR) has proven to be a valuable task in numerous NLP applications, including the evaluation of LLMs (Asaadi et al., 2019; Naseem et al., 2021). Determining the degree of relatedness in STR, however, remains a challenging task in computational semantics. That is because STR encompasses a broader range of commonalities beyond just meaning, including shared viewpoint, topic, and period, demanding a deeper understanding than semantic similarity alone (Asaadi et al., 2019; Abdalla et al., 2023). For example, consider reading these two sentences (*He heard the waves crashing gently*) and (*Making him feel calm and peaceful*). While humans easily recognize their strong relatedness and shared description of the same view (a beach scene), machines require advanced lexical and statistical methods to achieve the same level of understanding. STR techniques mainly come from four approaches: lexical similarity (Chen et al., 2018; Jarrar and Amayreh, 2019; Alhafi et al., 2019), semantic similarity (Hasan et al., 2020; Ghanem et al., 2023), deep learning (Zhang and Moldovan, 2019), and LLMs (Li et al., 2021).

Recently, Abdalla et al. (2023) introduced their STR-2022 dataset, which uses fine-grained scores ranging from 0 (least related) to 1 (completely related). Their dataset consists of 5,500 scored English sentence pairs. They framed the task as supervised regression, where they fine-tuned two language models, BERT-base (Kenton and Toutanova, 2019) and RoBERTa-base (Liu et al., 2019), and applied average pooling on top of the final embedding layer. Their testing of these models on the STR-2022 dataset yielded an average Spearman correlation of 0.82 for BERT-base and 0.83 for RoBERTa-base. On the other hand, their unsupervised experiments using Word2Vec (Mikolov et al., 2013) achieved a correlation score of 0.60, outperforming both BERT-base (0.58) and RoBERTa-base (0.48) by 0.02 and 0.12 points, respectively.

Asaadi et al. (2019) created the Bi-gram Semantic Relatedness Dataset (BiRD) for examining semantic composition. To avoid inconsistencies and biases from traditional 1-5 rating scales, they employed fine-grained scoring of bi-gram pairs (0-1) using the best-worst scaling (BWS) annotation technique (Kiritchenko and Mohammad, 2017). The dataset consists of 3,345 scored English term pairs. They utilised three models to generate word representations: GloVe (Pennington et al., 2014), FastText (Grave et al., 2018), and a word-context co-occurrence matrix (Turney et al., 2011). To calculate relatedness scores between pairs, they employed cosine similarity between the generated addition-pooled vectors. The FastText model achieved the highest performance with a Pearson correlation of 0.60.

The semantic relatedness between noun-pairs was studied using contextual similarity by Miller and Charles (1991). They attempted to understand distinctions between nouns in contextual discourse and how the similarity can be broader than just the meaning. Additional ideas could rely on extracting named entities (Liqreina et al., 2023; Jarrar et al., 2022) to measure the relatedness (Ghosh et al., 2023). However, the task evolved, leading to the creation of the up-to-date dataset presented by the SemRel-2024 shared task (Ousidhoum et al., 2024b). Their dataset annotation scores are at the level of sentence pairs. They shared baseline results for fourteen languages and dialects using Spearman correlation score. Since our focus is on Arabic, we have chosen its results to show. For example, their baseline is 0.42 for MSA in track B using multilingual BERT (mBERT) (Kenton and Toutanova, 2019), 0.60 for Algerian and 0.77 for Moroccan in track A using Label Agnostic BERT Sentence embeddings (LaBSE) (Feng et al., 2022). Specifically, their Algerian Arabic dataset offers 1,261 training and 583 test instances, Moroccan Arabic dataset includes 924 training and 425 test instances, and MSA Arabic dataset has 595 instances for testing.

Many efforts have been made to understand Arabic dialects, such as dialect identification, intent detection, and morphological annotations (Haff et al., 2022; Nayouf et al., 2023; Jarrar et al., 2023c, 2017, 2023a), but none studied STR between dialects.

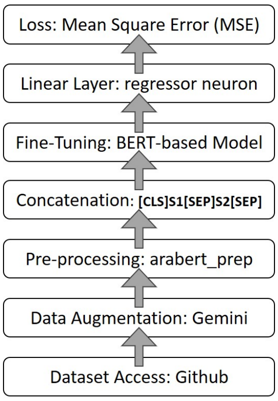

Loss: Mean Square Error (MSE)
Linear Layer: regressor neuron
Fine-Tuning: BERT-based Model
Concatenation: [CLS]S1[SEP]S2[SEP]
Pre-processing: arabert_prep
Data Augmentation: Gemini
Dataset Access: Github

Figure 1: BERT-based Supervised Architecture (A).

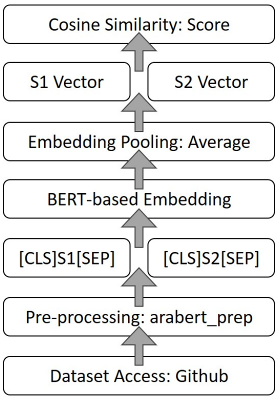


Figure 2: BERT-based Unsupervised Architecture (B).

3 System Overview

This section presents the techniques, datasets, and the augmentation we employed in tracks A and B.

3.1 Supervised Track (A)

Since the datasets use continuous scoring values, we tackled STR as a regression problem. We fine-tuned BERT with Mean Squared Error (MSE) objective. The model uses a regressor output layer, represented by a single neuron to predict the scores of the sentence-pairs. The data was pre-processed using the technique presented in (Antoun et al., 2020) to achieve standardized word forms. Before supplying the sentence pairs to the model, each was concatenated using the special tokens of the model input in this format:`[CLS]Sentence1[SEP]Sentence2[SEP]`. Figure 1 depicts our method architecture for the supervised track (A). Since we focused on the Algerian and Moroccan dialects in this track, we investigated various model parameters including learning rates, number of epochs, and pre-trained models to understand which model is better suited for each dialect. We found that both models, AraBERTv2 [2] and ArBERTv2 [3], best fits the Moroccan dialect more than Algerian. Nonetheless, we used same models for the Algerian dataset.

3.2 Unsupervised Track (B)

The STR using MSA is covered in track B (unsupervised learning), where training (or fine-tuning) on labeled data is not permitted. We employed cosine similarity (Reimers and Gurevych, 2019) as an unsupervised technique to calculate the sentence-pair scores. Figure 2 illustrates our architecture. We conducted initial experiments using the same aforementioned models, ArBERTv2 and AraBERTv2, for generating sentence representations. Various pooling options (CLS, average, max, and min) (Zhao et al., 2022) were applied on the final embedding layer in each (frozen) model, and found that AraBERTv2 with average-pooling is better suited for MSA in this track. The same data preprocessing used in track A is applied in B.

3.3 Datasets

The datasets provided by the SemRel-2024 shared task cover fourteen languages and dialects. In the paper, we used three Arabic datasets (Algerian, Moroccan, and MSA). Table 1 presents their data splits, including train, development, and testing. MSA has no labeled train data as it is included in Track B. However, for the other two dialects, we employed BERT-based models, that requires large train data (Bevilacqua et al., 2021).

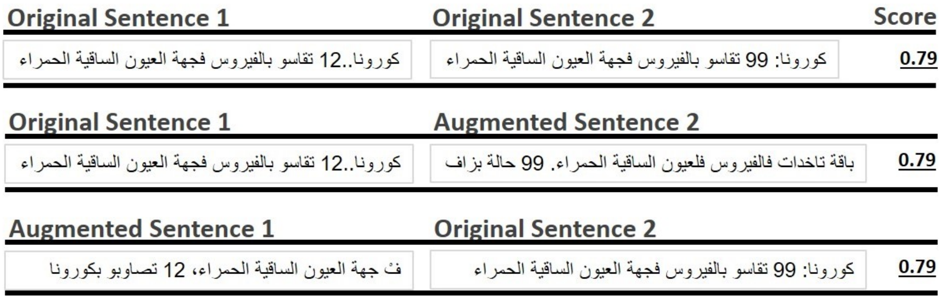
	MSA	Algerian	Moroccan
Train			
Original	–	1,261	924
Augments	–	–	757
Total	–	1,261	1,681
Dev.	32	97	70
Test	595	583	425

Table 1: The original and augmented datasets splits.

Different methods can be used for data augmentation, such as back-translation (Lin and Giambi, 2021) and generative models (Saidi et al., 2022). The back-translation technique was tested by (Malaysha et al., 2023) and showed minor improvement in performance. The availability of high-

[2] https://github.com/aub-mind/arabert

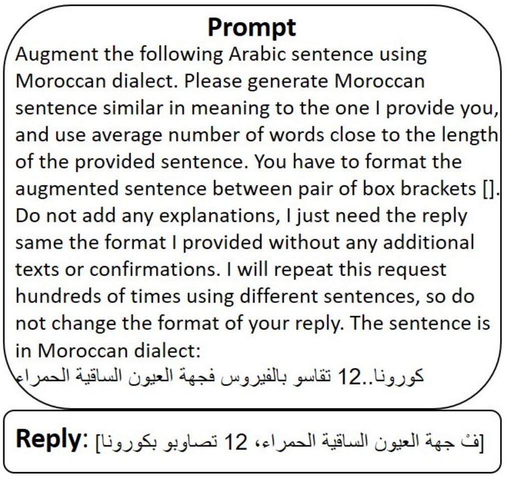
[3] https://huggingface.co/UBC-NLP/ARBERTv2



Original Sentence 1	Original Sentence 2	Score
كورونا..12 تقاسو بالفيروس فجهة العيون الساقية الحمراء	كورونا: 99 تقاسو بالفيروس فجهة العيون الساقية الحمراء	0.79
Original Sentence 1	**Augmented Sentence 2**	
كورونا..12 تقاسو بالفيروس فجهة العيون الساقية الحمراء	باقة تاخدات فالفيروس فلعيون الساقية الحمراء. 99 حالة بزاف	0.79
Augmented Sentence 1	**Original Sentence 2**	
فْ جهة العيون الساقية الحمراء، 12 تصاوبو بكورونا	كورونا: 99 تقاسو بالفيروس فجهة العيون الساقية الحمراء	0.79

Figure 3: Example of the augmented sentence-pairs.

quality generative models, such as ChatGPT [4] and Google Gemini, encouraged us to employ them in automatic augmentation. We employed in-context learning (Min et al., 2022) by prompting both models with the request depicted in Figure 4.



Prompt

Augment the following Arabic sentence using Moroccan dialect. Please generate Moroccan sentence similar in meaning to the one I provide you, and use average number of words close to the length of the provided sentence. You have to format the augmented sentence between pair of box brackets []. Do not add any explanations, I just need the reply same the format I provided without any additional texts or confirmations. I will repeat this request hundreds of times using different sentences, so do not change the format of your reply. The sentence is in Moroccan dialect:
كورونا..12 تقاسو بالفيروس فجهة العيون الساقية الحمراء

Reply: [فْ جهة العيون الساقية الحمراء، 12 تصاوبو بكورونا]

Figure 4: The prompt template employed for Gemini.

The initial manual reviews and tests for twenty prompts of Moroccan and Algerian sentences showed that both models are weak in Algerian comprehension. ChatGPT is also weak in the Moroccan, while Gemini demonstrated a high understanding of the Moroccan. Therefore, we decided to employ Gemini to augment the Moroccan train split. From every sentence-pair, we took each sentence and prompted it using the template in Figure 4. We mapped the augmented (new) sentence from the model with the other sentence in the same pair using the same score of the pair, as illustrated in Figure 3. By manually reviewing all the model replies, we found cases that were not valid (wrong content), and accordingly, we defined filters to exclude the not applicable data per the following rules:

- The model admits in the reply that it is just a language model and cannot fulfill the request. The model reply in such case has common format to rely on for the filter comparison.
- The case when the reply goes far from the original meaning. This option is achieved by manually reviewing the paraphrased contents.
- When the model rejects augmentation because the requested sentence contains information that breaks the model policy, i.e., talking about public figures or sensitive discussions. Similar to first rule, it has common reply format to automatically compare with.

Finally, after filtering the invalid augmentations, we reached 757 accepted sentences which we added to the Moroccan training set (See Table 1), reaching a total of 1,681 instances.

4 Experimental Setup

Our experiments fine-tuned two language models for Algerian and Moroccan, where we used the following pre-trained models: maubmindlab/bert-base-arabertv02 (Antoun et al., 2020) and UBC-NLP/ARBERTv2 (Abdul-Mageed et al., 2021). We employed the training data provided by the shared task, in addition to the data generated by our augmentation technique, when applied. The development data is excluded from either training or testing in the official evaluation phase, and testing is done on the shared task test set (See Table 1). The data pairs were concatenated using special tokens (`[CLS]` and `[SEP]`), as depicted in Figure 1, and digested by the models. The fine-tuning was done as a regression task using one neuron in the output layer, optimized using MSE as the loss function, and we used R-squared (Miles, 2005) to measure the improvement. The final hyper-parameters in the fine-tuning process were: 10 epochs for training, 4 epochs for early stopping, a batch size of 16,

[4] https://chat.openai.com/

Development Phase	**Track A**			**Track B**
	Algerian	Moroccan	Augmented Moroccan	MSA
ArBERTv2	0.55	0.82	0.88↑	0.42
AraBERTv2	0.69	0.84	0.79↓	0.58

Table 2: Our results on the development phase (i.e., on development split).

TEST Phase	**Track A**			**Track B**
	Algerian	Moroccan	Augmented Moroccan	MSA
Baseline (Ousidhoum et al., 2024a)	0.60	0.77	0.77	0.42
ArBERTv2	0.42↓	0.78↑	**0.83**↑	0.34↓
AraBERTv2	**0.53**↓	0.79↑	0.77↑	**0.49**↑

Table 3: The evaluation results on the test data. Our official ranked scores are in bold.

512 is the maximum sequence length, a learning rate of $2e^{-5}$, 50 evaluation steps, a seed of 42, and train (± augmented data) split.

In the experiments of B track for the MSA, no supervised fine-tuning is needed. Therefore, we neither used labeled data nor augmentation. We employed average-pooling on the embeddings of the sentence tokens from the final layer in each model. Then, we calculated the cosine similarity between the average embeddings of the sentences in each pair. This was done to estimate the fine-grained scores for the test (or development) data provided by the shared task. The shared task considers Spearman correlation score to evaluate the submitted predictions against their ground truth.

5 Results

Our approaches have achieved competitive ranks in the SemRel-2024 shared task. The official results of the tracks we participated in, as well as the baselines that were introduced by Ousidhoum et al. (2024a), are shown in Table 3. Additionally, our results on the development data are presented in Table 2. In the test evaluation, we ranked first in Track B for the MSA, with a Spearman correlation score of 0.49 using the AraBERTv2 model, outperforming the baseline by 0.07 points. However, ArBERTv2 did not perform well in Track B for MSA on both test and development splits. In contrast, ArBERTv2 achieved a high score in Track A for the Moroccan dialect when fine-tuned on both the train split and augmentation data, outperforming the baseline by 0.06 points on test split, ranking 5th among the submitted systems. Nonetheless, neither of the models, ArBERTv2 or AraBERTv2, surpassed the baseline for the Algerian dialect in Track A, where our rank is 12. Similarly, both models achieved low performance on the Algerian development split. It is possible that if we were able to augment the Algerian data as well, it could have performed better, similar to the improvement achieved in the Moroccan dataset. It is worth noting that AraBERTv2 outperformed both the baseline and ArBERTv2 on the original training data of the Moroccan dataset. However, its performance degraded on both test and development splits once the augmentation was included in the fine-tuning, unlike what happened with the ArBERTv2 model, on both splits. This could be due to the nature of the data utilized in the pre-training phase of the model. Due to the anisotropy problem (Baggetto and Fresno, 2022) inherent in BERT-based pre-trained models, we noted that computing cosine similarity directly between sentence representations is insufficient for discerning relatedness.

6 Conclusion

We presented our contributions to the SemRel-2024 shared task. We targeted three Arabic dialects covered by the shared task datasets, including MSA, Algerian, and Moroccan. Our approaches employed supervised and unsupervised techniques using commonly known language models, namely ArBERT and AraBERT. We augmented the training data using generative models, which enhanced the models' performance. Our system ranked first (MSA), fifth (Moroccan), and twelfth (Algerian) across the different tracks. We plan to augment additional data of Moroccan and Algerian using other models than what we used in this work. We will use the augmentations to experiment with both Arabic mono-dialect and cross-dialect fine-tuning.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. 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, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 782–796. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7088–7105. Association for Computational Linguistics.

Moustafa Al-Hajj and Mustafa Jarrar. 2021a. Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Moustafa Al-Hajj and Mustafa Jarrar. 2021b. Lu-bzu at semeval-2021 task 2: Word2vec and lemma2vec performance in arabic word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 748–755, Online. Association for Computational Linguistics.

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Shima Asaadi, Saif M. Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 505–516. Association for Computational Linguistics.

Alejandro Fuster Baggetto and Víctor Fresno. 2022. Is anisotropy really the cause of BERT embeddings not being semantic? In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4271–4281. Association for Computational Linguistics.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. SICK through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Lang. Resour. Evaluation*, 50(1):95–124.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338.

Zugang Chen, Jia Song, and Yaping Yang. 2018. An approach to measuring semantic relatedness of geographic terminologies using a thesaurus and lexical database sources. *ISPRS Int. J. Geo Inf.*, 7(3):98.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

Mirna El Ghosh, Nicolas Delestre, Jean-Philippe Kotowicz, Cecilia Zanni-Merk, and Habib Abdulrab. 2023. Reltopic: A graph-based semantic relatedness measure in topic ontologies and its applicability for topic labeling of old press articles. *Semantic Web*, 14(2):293–321.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomás Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. 9. Routledge.

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).

Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.

Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023a. Arbanking77: Intent detection neural model and a new dataset in modern and dialectical arabic. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 276–287. ACL.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammad Khalilia. 2023b. Salma: Arabic sense-annotated corpus and wsd benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023c. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 465–470. Association for Computational Linguistics.

Xiaotao Li, Shujuan You, and Wai Chen. 2021. Enhancing accuracy of semantic relatedness measurement by word single-meaning embeddings. *IEEE Access*, 9:117424–117433.

Guan-Ting Lin and Manuel Giambi. 2021. Context-gloss augmentation for improving word sense disambiguation. *arXiv preprint arXiv:2110.07174*, abs/2110.07174.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic fine-grained entity recognition. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 310–323. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sanad Malaysha, Mustafa Jarrar, and Mohammed Khalilia. 2023. Context-gloss augmentation for improving arabic target sense verification. In *Proceedings of the 12th Global Wordnet Conference, GWC 2023, University of the Basque Country, Donostia - San Sebastian, Basque Country, Spain, 23 - 27 January 2023*, pages 254–262. Global Wordnet Association.

Chenggang Mi and Shaoliang Xie. 2024. Language relatedness evaluation for multilingual neural machine translation. *Neurocomputing*, 570:127115.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jeremy Miles. 2005. R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science*.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(5):74:1–74:35.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 12–23. ACL.

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. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Ahmed Sabir, Francesc Moreno, and Lluís Padró. 2019. Semantic relatedness based re-ranker for text spotting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3449–3455. Association for Computational Linguistics.

Rakia Saidi, Fethi Jarray, Jeongwoo Jay Kang, and Didier Schwab. 2022. GPT-2 contextual data augmentation for word sense disambiguation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, PACLIC 2022, Manila, Philippines, October 20-22, 2022*, pages 455–462. De La Salle University.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2014. Text relatedness based on a word thesaurus. *CoRR*, abs/1401.5699.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 680–690. ACL.

Linrui Zhang and Dan Moldovan. 2019. Multi-task learning for semantic relatedness and textual entailment. *Journal of Software Engineering and Applications*, 12(6):199–214.

Shuai Zhao, Tianyu Zhang, Man Hu, Wen Chang, and Fucheng You. 2022. AP-BERT: enhanced pretrained model through average pooling. *Appl. Intell.*, 52(14):15929–15937.