L3IRIT at SemEval-2026 Task 4: Learning Narrative Similarity from Aligned Film Plot Summaries

Ahmed Hamdi[1], Emanuela Boros[2], José G. Moreno[1], Adam Jatowt[3], Georgeta Bordea[1], Carlos-Emiliano González-Gallardo[4], Antoine Doucet[2,5],
[1]IRIT, University of Toulouse,
[2]L3i, University of La Rochelle,
[3]DiSC, University of Innsbruck,
[4]LIFAT, University of Tours,
[5]FRI, University of Ljubljana
Correspondence: ahmed.hamdi@irit.fr

Abstract

This paper presents the participation of the L3IRIT team in SemEval Task 4.The team is a joint research group working on narrative extraction from historical text, led by the IRIT laboratory (University of Toulouse) and the L3i laboratory (University of La Rochelle). Our participation is grounded in the construction of a novel bilingual resource extracted from Wikipedia by automatically aligning film plots. Leveraging this dataset, we train embedding models using contrastive learning objectives to capture higher-level narrative structures more effectively. The resulting resource goes beyond surface-level lexical overlap, providing supervision for narrative similarity without manual annotation. In addition, we introduce a named-entity masking strategy designed to promote narrative abstraction and reduce superficial entity-based matching. Overall, our approach aims to support representation learning that captures structural and event-level similarities across stories in different languages more effectively. Our system ranked in 24 of the 44 scoreboards for Task A and 20 of the 27 scoreboards for Task B, achieving accuracies of 65.75% and 61.00%, respectively.

1 Introduction

Recent advances in representation learning, particularly embedding models based on pretrained language models such as Devlin et al. (2019) and Liu et al. (2019) have substantially improved semantic similarity estimation. Sentence transformers and their variants (Reimers and Gurevych, 2019; Thakur et al., 2021) generate dense representations that capture contextual meaning beyond simple lexical overlap and have become standard tools for semantic textual similarity and paraphrase detection (Artetxe and Schwenk, 2019). However, most existing benchmarks and training resources operate at the sentence level and focus on semantically related texts. They rarely address higher-level narrative alignment, where similarity depends not only on semantic proximity but also on shared event sequences, character roles, and structural correspondences (Cer et al., 2017; Agirre et al., 2012; Conneau et al., 2017; Hill et al., 2016; Bjerva and Östling, 2017). As a result, current models often remain sensitive to superficial variations such as character names, locations, or stylistic reformulations, while failing to capture deeper narrative equivalence (McCoy et al., 2019; Chambers and Jurafsky, 2008; Goyal et al., 2010; Elson, 2012). Based on vanilla sentence transformers, two texts describing the same storyline may therefore be judged dissimilar if they differ in surface realization. In this work, we address the challenging task (Hatzel et al., 2026) of learning to represent narrative similarity in English stories, enabling models to recognize shared narratives despite variations in surface forms.

We address this limitation by introducing a novel bilingual resource designed specifically for learning narrative-level representations. The dataset is constructed from aligned film plot summaries automatically extracted from English Wikipedia and English translations from Portuguese Wikipedia. Film plots constitute a structurally coherent narrative domain in which recurring story patterns and event structures naturally emerge, making them suitable for studying narrative similarity at scale.

We collect plot summaries in both languages and exploit cross-lingual alignment to construct paired narratives that describe the same underlying story. Although these plots refer to identical films, they differ in level of detail, narrative emphasis, discourse organization, and lexical realization. We treat such aligned summaries as positive examples of narrative similarity, assuming that translations preserve core event structures while introducing controlled surface variation. This strategy enables large-scale supervision without manual annotation.

Building on this resource, we train embedding

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1853–1859

models using contrastive objectives tailored to narrative alignment. To further encourage abstraction beyond surface identifiers, we introduce a named-entity masking strategy that reduces reliance on character names and location markers during training. This approach promotes representations that emphasize event structure and narrative progression rather than lexical cues. Through this contribution, we aim to advance representation learning for narrative texts and provide a resource that facilitates research on cross-lingual narrative similarity and narrative representation learning. Our work is similar to Hatzel and Biemann (2024) who align film summaries across multiple languages leveraging cross-lingual correspondences between descriptions of the same underlying stories. However, unlike this prior work, which primarily focuses on retrieval of summaries based on shared story identity and often relies heavily on named entities as retrieval signals, our objective is to learn representations that explicitly abstract away from lexical overlap and entity memorization, rather than optimizing for an information retrieval purpose.

Our participation in the task demonstrates that fine-tuning models on a dataset of film plot summaries improves baseline models' performance in narrative detection and that an entity-masking strategy yields additional gains. Our system achieved 65.75% in Track A (Narrative Story Similarity) and 61% in Track B (Narrative Representation Learning) on the test set, ranking 24^{th} out of 44 and 20^{th} out of 27 participating teams, respectively (Hatzel et al., 2026). All experiments were conducted using open-source sentence transformers, without the use of commercial large language models (LLMs). Our code and trained models are publicly available at `https://github.com/ahHamdi/narrative-similarity-for-SemEval-2026-task-4`.

The remainder of the paper is organized as follows: Section 2 presents the task and the data. We then describe the construction of the multilingual narrative resource and the experimental setup in Section 3. Section 4 reports results on narrative similarity ranking, followed by conclusions and future research directions in Section 5.

2 Background

The shared task Narrative Story Similarity and Narrative Representation Learning (Hatzel et al., 2026) focuses on identifying narratively similar stories by modelling similarity beyond surface-level lexical overlap. Narrative similarity is defined along three core and complementary dimensions: the abstract theme, the course of action, and the story's outcomes. The abstract theme refers to the underlying ideas, motives, or conceptual backbone that structure a narrative, such as revenge, redemption, or personal transformation. The course of action captures the sequence of central events, including conflicts, turning points, and causal developments that shape the story's progression. Outcomes correspond to the final states or consequences resulting from these events. Together, these three components define narrative similarity as a structural and conceptual relation rather than a purely lexical or topical one.

Track A operationalizes this notion of similarity as a relative ranking task. Each instance consists of a triple composed of an anchor story and two candidate stories. The system must determine which of the two candidates is narratively closer to the anchor. This formulation evaluates the model's ability to discriminate between competing narrative alignments, even when lexical cues may be misleading. Two stories may share surface entities or settings while diverging in narrative arc, whereas another pair may differ lexically yet exhibit strong structural correspondence in their thematic development, event sequence, and resolution. The task, therefore, emphasizes comparative narrative judgment rather than absolute similarity scoring.

Track B instead focuses on narrative representation learning. Participants are required to produce a vector representation for each individual story such that the cosine similarity between the embeddings reflects the underlying narrative similarity. The evaluation is performed by comparing similarity relations derived from the embedding space against the triple-wise similarity judgments provided by the organizers. In this setting, the quality of a system is measured by how well its representation space encodes abstract thematic alignment, event structure, and outcome correspondence. Unlike Track A, which evaluates discrete ranking decisions, Track B assesses the coherence and generalization capacity of learned narrative embeddings.

We participated in both tracks to comprehensively evaluate our approach. The dataset provided by the organizers consists of English story summaries. The final evaluation is conducted on a test set comprising 400 triplet instances (an anchor with two candidates) for Track A and 849 individual stories for Track B. In this work, we rely

exclusively on the development dataset for experimentation and validation, using it to identify the best-performing model for submission to the test set. For Track A, the development set comprises 200 labelled triplet instances, which are also available as individual story items, thereby enabling evaluation in the Track B setting. We use the development dataset solely as an evaluation benchmark and do not perform supervised training on it. Instead, our models are trained on an external bilingual resource, and the official development data serves to assess the transferability of the learned narrative representations. This setup allows us to evaluate whether large-scale cross-lingual narrative pairing provides embeddings that align with curated human judgments of narrative similarity.

3 System Overview

This section presents the overall architecture of our system and the methodological approach adopted for the shared task. We build a narrative dataset and train embedding models using contrastive learning objectives, combined with entity masking strategies to promote structural abstraction.

3.1 Data Sourcing

We constructed our bilingual dataset from film plot summaries extracted from Wikipedia. Films constitute a large-scale and structurally coherent narrative domain in which recurring story patterns and well-documented plot structures naturally emerge. This makes them particularly suitable for studying narrative similarity at scale.

The list of candidate films was derived from curated index pages grouped under Wikipedia[1]. For each selected film, we retrieved the corresponding Wikipedia article content using the Hugging Face Wikimedia Wikipedia dataset[2], which provides structured Wikipedia dumps suitable for large-scale querying and processing. From these articles, we extracted plot descriptions in English (EN) and Portuguese (PT). Narrative-focused sections correspond to content labeled *plot* or *synopsis*, depending on language conventions.

The plots were aligned across languages using Wikipedia interlanguage links and title matching. This alignment process pairs English and Portuguese summaries that describe the same film and therefore the same underlying narrative. Although these paired plots refer to identical stories, they naturally differ in several respects: they may vary in levels of detail, narrative emphasis, discourse organization, and lexical realization. Some are direct translations, while others are independently written summaries of the same source material. These differences are not treated as noise but as meaningful surface variation that models must learn to abstract away from in order to capture deeper narrative equivalence.

To further increase comparability and enable controlled narrative pairing within a shared representational space, we automatically translated Portuguese plot summaries into English using the Google Translation API[3]. We then constructed aligned narrative pairs by treating each original English plot and its translated Portuguese counterpart as narratively equivalent texts. This design yields pairs that share a common narrative backbone while exhibiting controlled lexical and syntactic variation arising from cross-lingual differences and the translation process itself.

This resource enables us to consider narrative similarity as a learning objective. By treating aligned pairs as positive examples, we train embedding models using contrastive objectives that pull representations of narratively equivalent texts closer together while pushing apart representations of different stories. This approach provides supervision for narrative-level abstraction: models learn to recognise that two texts describe the same story even when they differ in surface form, expression style, or language of origin. The resulting representations are thus encouraged to capture underlying event sequences, character roles, and structural narrative patterns rather than relying on superficial lexical overlap.

The resulting dataset consists of more than $7,000$ aligned English-Portuguese narrative pairs, with English plots averaging over 500 tokens per summary, reflecting the richness and structural depth of the narratives. Our source code and the generated resource are available via this link[4].

3.2 Entity Masking for Narrative Abstraction

Narrative texts, such as movie plots, are typically rich in named entities, including character names, locations, and organizations. While such entities

[1] https://en.wikipedia.org/wiki/Lists_of_films

[2] https://huggingface.co/datasets/wikimedia/wikipedia

[3] https://cloud.google.com/translate

[4] https://github.com/ahHamdi/narrative-similarity-for-SemEval-2026-task-4

are integral to storytelling, they may introduce lexical shortcuts that allow models to infer similarity based on shared identifiers rather than deeper narrative structure. For instance, two plot variants of the same film may exhibit high lexical overlap primarily due to repeated character names, even when describing different stages of the narrative.

To mitigate this effect and encourage abstraction, we introduce a named entity masking strategy. Entities corresponding to persons, locations, organizations, and geopolitical entities are detected using the off-the-shelf named entity recognition system spaCy[5]. Each detected entity is replaced with its corresponding entity type label (e.g., PERSON, GPE, DATE). This transformation suppresses entity identity while preserving syntactic structure and event relations.

The underlying hypothesis is that many stories share comparable thematic arcs, event progressions, and outcomes while differing primarily in protagonists or settings. Entity masking, therefore, acts as an explicit inductive bias toward narrative abstraction, encouraging embedding models to capture structural and event-level correspondences rather than memorizing entity names.

3.3 Contrastive Training Objective

To learn narrative-aware representations from the generated resource, we adopt a contrastive learning framework. Given an aligned pair of plot summaries describing the same underlying story, we treat the two texts as a positive pair, while other plots within the batch serve as implicit negatives. This formulation encourages the model to maximize similarity between narratively equivalent variants while separating unrelated stories in the embedding space.

Let f_θ denote the embedding function parameterized by θ, mapping a plot summary x to a dense vector representation $\mathbf{h} = f_\theta(x) \in \mathbb{R}^d$. Representations are ℓ_2-normalized, and the similarity between two plots x_i and x_j is computed using cosine similarity:

$$\mathrm{sim}(x_i, x_j) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}.$$

For each positive pair (x_i, x_i^+), the model is trained using an InfoNCE-style objective:

$$\mathcal{L}_i = -\log \frac{\exp(\mathrm{sim}(x_i, x_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(x_i, x_j)/\tau)},$$

where τ is a temperature hyperparameter and N is the batch size. All other batch examples act as negative candidates. The loss is averaged across all positive pairs. This objective aligns naturally with the narrative similarity task. By bringing together cross-lingual plot variants that share the same abstract theme, event progression, and outcome, the model is encouraged to encode higher-level narrative structure rather than surface lexical overlap. Combined with the entity masking strategy, contrastive training promotes abstraction from specific character names and locations while preserving event-level coherence.

The resulting embedding space is therefore structured such that cosine similarity reflects narrative proximity. This property directly supports both Track A, where relative similarity comparisons are required, and Track B, where the geometric consistency of narrative representations is evaluated.

4 Experiments and Results

We evaluate the effectiveness of our bilingual narrative dataset for learning cross-lingual narrative similarity by fine-tuning three lightweight sentence transformer models: `all-MiniLM-L6-v2`, `all-MiniLM-L12-v2`, and `all-mpnet-base-v2`. These models were selected for their balance between computational efficiency and strong performance on semantic similarity tasks.

We conduct fine-tuning using contrastive learning on our aligned English-Portuguese narrative pairs. We compare two training configurations:

- **Base fine-tuning**: Models are fine-tuned on the original aligned plot summaries, treating each English-Portuguese pair as a positive example of narrative equivalence. We use a batch size of 8 pairs and train for 3 epochs. We employ the AdamW optimizer with a learning rate of 3×10^{-5}. The maximum sequence length is set to 128 tokens to focus on the core narrative content.
- **Entity-masked fine-tuning**: Models are fine-tuned on the same data but with named entities (character names, locations, dates) masked during training. This strategy encourages abstraction away from surface identifiers and promotes focus on narrative structure.

[5] https://spacy.io/

The fine-tuned models are used for both Track A and Track B. All models are evaluated on the development set, and performance is reported using the standard metric, accuracy. Table 1 presents the evaluation results for all model configurations. For each model, we report performance after standard fine-tuning and after entity masking. For evaluation purposes, results are reported primarily on Track A, under the assumption that the best-performing system on Track A is also the best-performing system for Track B.

Table 1: Accuracy (acc.), kurtosis (kurt.), and skewness (skew.) results for Track A on the narrative similarity development set.

Model	Train	Acc.	Kurt.	Skew.
MiniLM-L6	None - Vanilla model	0.550	1.125	0.292
	Standard Finetuning	0.610	0.082	-0.031
	+ Entity Masking	0.620	-0.216	-0.022
MiniLM-L12	None - Vanilla model	0.565	0.294	0.103
	Standard Finetuning	0.600	-0.050	-0.070
	+ Entity Masking	0.610	-0.161	0.012
mpnet-base	None - Vanilla model	0.625	-0.027	-0.015
	Standard Finetuning	0.660	-0.159	0.014
	+ Entity Masking	**0.680**	-0.192	-0.103

Table 1 shows consistent improvements across all models through fine-tuning on our bilingual narrative dataset, with gains from 3.5 to 6.0 percentage points over baselines. This confirms that our resource provides effective supervision for learning narrative-level representations. `mpnet-base` achieves the best performance, reaching 0.68 accuracy with entity masking, outperforming the MiniLM variants (0.62 and 0.61). This suggests that its larger capacity is better suited for capturing structural narrative correspondences. Entity masking yields small but consistent gains (1.0–2.0 points) across all models, supporting our hypothesis that reducing reliance on surface identifiers improves structural representations.These improvements generalize to the test set, where our system, based on the finetuned mpnet with entity masking, ranks in 24 out of 44 scoreboards for track A and 20 out of 27 for track B, achieving accuracies of 65.75% and 61.00%, respectively.

From the development set, we constructed triplets (a, p, n) by selecting the anchor text a and assigning the text annotated as more similar to the anchor to the positive example p, while the remaining text is used as the negative example n.

To evaluate the impact of fine-tuning and entity masking, we compute the similarity margin

$$\Delta = \mathrm{sim}(a, p) - \mathrm{sim}(a, n),$$

for each triplet, where a denotes the anchor, p the positive sample, and n the negative sample. A positive margin indicates correct ranking, whereas negative values correspond to triplet violations.

Figure 1 presents the distribution of Δ for three encoder families (MiniLM-L6, MiniLM-L12, and MPNet), comparing baseline models, fine-tuned models and variants with entity masking.

Across all three model families, fine-tuning consistently shifts the margin distribution to the right, indicating improved discrimination between positive and negative samples. The reduction of mass below $\Delta = 0$ reflects a decrease in triplet violations and directly explains the observed gain. Moreover, the entity-masked variant further amplifies this effect with a distribution more concentrated in the positive region and larger mean margins. This suggests that masking entities during fine-tuning reduces reliance on surface-level lexical overlap and promotes semantic understanding, leading to stronger anchor–positive alignment. Importantly, this pattern holds consistently across MiniLM-L6, MiniLM-L12, and MPNet, demonstrating that the benefit of entity masking is architecture-agnostic rather than model-specific. These observations are supported by the kurtosis and skewness values in Table 1. Kurtosis decreases across fine-tuning strategies, indicating fewer extreme deviations and more stable predictions. Skewness also moves closer to zero, suggesting a more symmetric distribution. Overall, these results indicate that fine-tuning, especially with entity masking, leads to higher accuracy and prediction stability.

5 Conclusion

For the participation of our team L3IRIT in SemEval-2026 Task 4, we proposed an English-Portuguese dataset constructed from aligned film plot summaries to support learning cross-lingual narrative similarity. We fine-tuned three lightweight sentence transformer models using contrastive objectives and evaluated a named-entity masking strategy to encourage abstraction beyond surface identifiers. Our results show consistent improvements across all models, with `mpnet-base` achieving the best performance. The entity masking strategy yielded small but consistent gains, confirming that reducing reliance on character names and locations promotes deeper narrative structure.

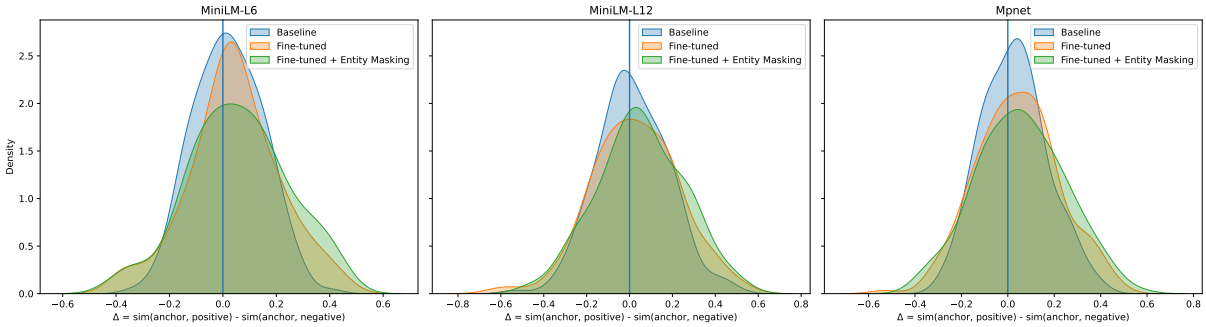

Figure 1: Distribution of similarity margins $\Delta = \text{sim}(a, p) - \text{sim}(a, n)$ across three model families. Fine-tuning shifts the margin distribution toward positive values, while entity masking further increases the separation between positive and negative samples.

Acknowledgments

This work has been co-funded by the French national research agency (ANR) through projects ANR-25-CE38-6695 (MILL-EHNAS) and ANR-22-CPJ2-0107-01 (CPJ), as well as by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 — AI4DH. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The authors would like to warmly thank Pierre Labardin and Angelo Riva, Full Professors at IAE La Rochelle and INSEEC Grande École de Paris, respectively, who lead the MILL-EHNAS project.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. in* sem 2012: The first joint conference on lexical and computational semantics–volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montréal, QC, Canada*, pages 7–8.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Johannes Bjerva and Robert Östling. 2017. Cross-lingual learning of semantic textual similarity with multilingual word representations. In *The 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, page 211. Linkoping UP.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 1–14.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

David K Elson. 2012. *Modeling narrative discourse*. Columbia University.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA. Association for Computational Linguistics.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2026 Task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024. Tell me again! a large-scale dataset of multiple summaries for the same story. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15732–15741.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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)*, pages 3982–3992.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 296–310.