Coherent or Not? Stressing a Neural Language Model for Discourse Coherence in Multiple Languages

Dominique Brunato[⋆], **Felice Dell'Orletta**[⋆], **Irene Dini**[⋆◇], **Andrea Amelio Ravelli** [⋆•]
[⋆]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – www.italianlp.it
[◇] University of Pisa
[•] University of Bologna
{dominique.brunato, felice.dellorletta, irene.dini}@ilc.cnr.it,
andreaamelio.ravelli@unibo.it

Abstract

In this study, we investigate the capability of a Neural Language Model (NLM) to distinguish between coherent and incoherent text, where the latter has been artificially created to gradually undermine local coherence within text. While previous research on coherence assessment using NLMs has primarily focused on English, we extend our investigation to multiple languages. We employ a consistent evaluation framework to compare the performance of monolingual and multilingual models in both in-domain and out-domain settings. Additionally, we explore the model's performance in a cross-language scenario.

1 Introduction

Coherence is a fundamental aspect of a well-organized text and it can be defined as "a semantic property of discourse, based on the interpretation of each individual sentence relative to the interpretation of other sentences" (Van Dijk, 1977). In order to be fully coherent, a discourse must exhibit both a local and a global coherence, where the former concerns mainly the relationships between adjacent or nearby sentences whereas the latter focuses on the discourse-level relations connecting remote sentences. Modeling discourse coherence has a long history in the NLP community, particularly in the "pre-deep-learning" era, where a great deal of studies was inspired to the *Centering Theory* framework (Grosz et al., 1995), such as the popular entity-grid approach for measuring local coherence (Barzilay and Lapata, 2008).

The long-standing interest for coherence modeling has been also motivated by the large variety of downstream applications which can benefit by measuring coherence in text, such as automatic essay scoring in language learning scenarios (Lai and Tetreault, 2018; Mesgar and Strube, 2018), language assessment in clinical settings (Elvevåg et al., 2007; Iter et al., 2018), readability assessment (Muangkammuen et al., 2020). A further emerging scenario, which is closer to our study, is related to research on the interpretability of modern deep neural networks. In this respect, while the majority of existing tasks and benchmarks on which NLMs are evaluated focuses on properties acquired from stand-alone sentences, their ability to model discourse and pragmatic phenomena is still unclear. Few exceptions are represented by recent works such as (Shen et al., 2021; Chen et al., 2019; Farag et al., 2020), which introduced dedicated test suites aimed at measuring if neural sentence representations show sensitivity to discourse phenomena spanning across sentences. Our paper intends to provide a novel contribution to the current body of literature by investigating whether and to what extent NLMs in multiple languages are able to distinguish a coherent piece of text from an incoherent one, where the latter has been artificially created to undermine local coherence within text, at gradual levels of difficulty. While all previous work on coherence assessment using NLMs has been focused on English, we probed these models for multiple languages using the same evaluation framework and compared the performance achieved by monolingual and multilingual models both in a in-domain and out-domain setting, as well as in a cross-language scenario.

Our Contributions This paper makes the following contributions: i) we devised a new task to model discourse coherence understanding; ii) we compiled two new multilingual datasets (freely available) representative of two different domains and levels of complexity, containing coherent and incoherent passages (artifically manipulated); iii) we assessed how a multilingual NLM, XLM-RoBERTa-base, performs over the task and compared the performance against the model without pretraining in order to measure the impact of pretraining on the task at hand; iv) we evaluated the task performance in a cross-domain and cross-

Findings of the Association for Computational Linguistics: ACL 2023, pages 10690–10700

Prompt	Target	Coherence
In 1998, an intense flare was observed. The star has also been a target of plans for interstellar travel such as Project Daedalus. In 2005, astronomers using data from the Green Bank Telescope discovered a superbubble so large that it extends beyond the plane of the galaxy.	It is called the Ophiuchus Superbubble.	✓
What do they do? Well, let's first check and make sure they're really amnesiac. We ask these amnesiac patients to tell us which one they own, which one they chose last time, which one is theirs.	Here's what normal controls do: they synthesize happiness.	✗

Table 1: Examples of prompt-target pairs: coherent the first; incoherent the latter.

lingual setting to test the generalization abilities of the model.

2 Coherence Evaluation Framework

We formulated the task of coherence modeling as a binary classification problem, that is: given a short piece of text (hereafter referred to as *prompt*) along with an individual sentence (the *target*), the model is asked to predict whether the target is contiguous or not, thus joining it to the prompt gives out a coherent or incoherent text. See Table 1 for reference. More specifically, we designed two task configurations, namely *forward* and *backward*. In the *forward* configuration, the model is asked to assess if the target follows the closing sentence of the prompt, whereas in the *backward* one if it comes before the initial sentence of the prompt. Regardless the direction, the negative target was always selected as either occurring in the same document of the prompt or randomly chosen from a different document. When selecting the target from the same document, we specifically chose it as the 5th or 10th sentence preceding the first or following the last sentence of the prompt.

By systematically manipulating the distance from the prompt we had intended to introduce incremental degrees of complexity in approaching the task, assuming that candidates closer to the prompt would pose a higher likelihood of being misleading.

We tested our approach on the following languages: English, French, Italian, Portuguese, and Spanish.

2.1 Dataset construction

For each language we built two distinct datasets, which were chosen as representative of both written and spoken modalities: on one side, we exploit the well known and (ab)used Wikipedia data; on the other, we relied on TED talks transcriptions. The latter can be seen as a *middle* modality in between written and spoken. Indeed, even if public speeches are performed orally, they often derive from written scripts, and they are prepared and rehearsed in advance. It derives that these communication events lack the typical *spontaneity* (Chafe, 1994) that characterize everyday oral communications and they do not contain phenomena such as false starts, retracting, and on-line discourse generation, thus they cannot be considered as natural spoken language examples. Nevertheless, TED-style talks represent a different domain with respect to Wikipedia, and in general to 'standard' written language, thus we included these transcriptions to test NLMs in a slightly more complex scenario.

As anticipated, the data source used to build the *written* section of the dataset is Wikipedia. Texts have been automatically extracted from the dumps and cleaned using Wikiextractor (Attardi, 2015).

The *spoken* section of the dataset has been derived from two sources, both collecting TED talks, i.e. the multilingual TEDx Dataset (Salesky et al., 2021),[1] and the TED2020 Dataset (Reimers and Gurevych, 2020)[2]. The latter has been used to include English data, that are not present in the TEDx Dataset. We discarded aligned translations, in order to collect exclusively original monolingual data.

To ensure consistent analysis for coherence assessment, we extracted passages consisting of four consecutive sentences from each text, considering them as our unit of analysis. As regards the *written* dataset, we utilized the existing paragraph

[1] https://www.openslr.org/100
[2] https://opus.nlpl.eu/TED2020.php

segmentation in Wikipedia to select four-sentence paragraphs. For the *spoken* one, given that TED speeches lack such an internal structure, we split all the transcripts into passages of four sentences.

To meet the requirement of identifying negative targets within a maximum distance of 10 sentences from the beginning or end of the prompt, we only retained prompts for which it was possible to retrieve such targets in both directions. Once a prompt was paired with a correct target, it was excluded from being a source for extracting negative items, and vice versa. It is worth noting that the positive items remained the same across all experiments, while the negative items, which shared the prompt but varied in the target sentence, were unique to each experimental variant.

Following these constrains, we ended up with a train-test dataset splits respectively of 8000 and 800 prompt-target pairs for each language, domain and configuration.[3] An example for each configuration in the dataset can be found in Appendix A.

2.2 Experimental settings

To evaluate our coherence assessment framework we devised three main sets of experiments. In the first one (*in-domain /in-language*), we examined the ability of a multilingual NLM to comprehend local coherence for each language and domain. To determine the impact of the linguistic knowledge acquired by the model during pretraining on the task, we compared its performance with a baseline model that lacked pretraining.

Through the second set of experiments we proceeded to evaluate the generalization abilities of the multilingual model in a *cross-language* scenario. Thus, the model was trained on one language and tested against all others. We compared the scores with the ones obtained by the same multilingual model trained on all languages simultaneously, and with the ones obtained by a monolingual model trained only in the corresponding language.

The last set of experiments (*cross-domain*) aimed at assessing whether and to what extent the model is able to learn information about coherence that can be generalized across datasets: for each language, we thus computed the performances of the multilingual model trained on one domain and tested on the other.

As regards the multilingual model used in the experiments, we relied on XLM-RoBERTa-base (Ruder et al., 2019), a multilingual version of RoBERTa-base (Liu et al., 2019) pretrained on 2.5TB of data containing 100 languages (including those under examination). The model consists of 12 layers with 12 attention heads. The monolingual model was chosen so as to be more similar as possible to xlm-RoBERTa-base and available within the Huggingface released models. Accordingly, we used the original RoBERTa-base (Liu et al., 2019) for English, the BERTIN version of RoBERTa (la Rosa et al., 2022) for Spanish, CamemBERT (Martin et al., 2020) for French, GilBERTo[4] for Italian. As regards Portuguese, since a reference version of this model is not available for this language we chose the RoBERTa model most used by the community [5]. For all settings, the passages have been fed to the examined model by simply concatenating the target sentence to the prompt without using special characters for separation. All experiments were executed using the Huggingface library[6] and the models were trained for 10 epochs since experiments on more epochs showed no improvements in terms of convergence. The remaining training hyper-parameters were set to their default values as specified by the Hugging Face framework, with the exception of the learning rate, which was set to 5e-6.

3 Results

Figure 1 reports the results of the multilingual model in the *in-domain* setting across languages for both *written* (top) and *spoken* (bottom) domain. As a general remark, we observe that the *baseline* model reports chance-level performance or even below across all task configurations and languages. This suggests that the knowledge acquired in the pre-training phase enables the model to capture information that is involved in local coherence. Such an impact is particularly beneficial when the negative target is sourced from a different text, especially evident in the Wikipedia data, where the pretrained model achieves an average F-score of 0.94 across all languages, compared to the 0.83 achieved on the TED data.

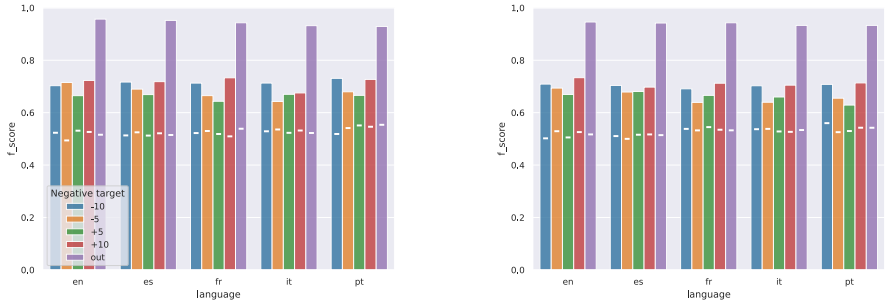
Conversely, although still performing better than the baseline, the model's performance significantly decreases when the negative target belongs to the

[3] All datasets are available at the following link: https://github.com/irenedini/coherence_dataset

[4] https://github.com/idb-ita/GilBERTo

[5] https://huggingface.co/josu/roberta-pt-br

[6] https://huggingface.co/



(a) Wikipedia - forward configuration

(b) Wikipedia - backward configuration

(c) TED - forward configuration

(d) TED - backward configuration

Figure 1: Summary of the in-domain classification scores of the multilingual model across languages. Columns represent F-score obtained in the different classification settings. The colors indicate the distance of the negative target from the prompt: *5, 10* sentences far from the prompt ('-' preceding/'+' following the prompt); *out*, the negative target belongs to a different document. The white dash within each column represents the score of the *baseline* model, which is the multilingual model without pretraining.

same document of the prompt. This effect becomes more pronounced as the target gets closer to the prompt (either preceding or following it). This suggests that the model tends to rely more on explicit lexical clues to detect incoherent passages and may be more confounded when the target and the prompt share the same topic. This observation is particularly relevant in TED speeches, where clear topic distinctions are less prominent, and the discourse structure is less defined compared to encyclopedic written articles. In this case, the "easiest" scenario of out-domain negative targets becomes more challenging, indicating that the model struggles to grasp coherence-related cues beyond lexical or semantic ones.

Taken overall, these results highlight that the out-document negative items are extremely easy to detect, whereas similar scores are obtained despite the configuration and prompt-pair distance. Based on this, we decided to conduct the cross-language and cross-domain experiments exclusively using the forward configuration, where the negative targets correspond the 10th sentence following the prompt.

	EN	ES	FR	IT	PT
EN	0.75	0.67	0.70	0.69	0.73
ES	0.72	**0.69**	0.71	0.68	0.72
FR	0.71	0.67	0.71	0.69	0.70
IT	0.71	0.68	0.69	**0.71**	0.71
PT	0.74	0.65	0.73	0.67	**0.74**
ALL	0.75	**0.69**	**0.74**	**0.71**	**0.74**
MONO	**0.76**	0.63	**0.74**	**0.71**	0.59

	EN	ES	FR	IT	PT
EN	0.69	0.67	0.67	0.67	0.66
ES	0.67	0.67	0.66	0.66	0.64
FR	0.68	0.67	0.66	0.66	0.69
IT	0.69	0.68	0.68	0.68	0.68
PT	0.68	0.66	0.66	0.68	0.68
ALL	0.72	**0.72**	**0.71**	**0.70**	**0.71**
MONO	**0.74**	0.70	**0.71**	0.67	0.53

Table 2: Cross-language F-score on Wiki (top) and TED (bottom). Row and column labels indicate respectively the language on which the model was fine-tuned and tested. ALL means the joint fine-tuning and MONO reports the results of the monolingual RoBERTa model for each language.

Tables 2 provides a summary of the results of the cross-language experiments. As we can see, the best overall scores are obtained by the multilingual model fine-tuned with all languages (row ALL in both Tables), especially for the TED dataset. As expected, training the model in a language dif-

ferent from the target language leads to slightly lower performance, although the differences are not dramatic. Interestingly, the monolingual model performs comparably to the multilingual model, except for Portuguese.

	EN	ES	FR	IT	PT
TED-WIKI	0.67	0.63	0.70	0.63	0.67
WIKI-WIKI	0.72	0.72	0.73	0.68	0.73
WIKI-TED	0.63	0.60	0.52	0.64	0.55
TED-TED	0.69	0.67	0.66	0.68	0.68

Table 3: Model performances in the cross-dataset experiments.

Shifting our focus to the cross-domain classification (Table 3), we observe a considerable decrease in performances for models fine-tuned on one domain and tested on the other, as anticipated. This holds especially when the model is tested on the Ted datasets. We can attribute this phenomenon to the less structured nature of TED speeches compared to the Wikipedia texts, but also to the fact that Wikipedia texts are part of the base training of the models. This effect is particularly appreciable if we look at the performances on French or Portuguese languages, but less marked in on Italian data.

4 Conclusion

In this study we carried out a comprehensive series of experiments to evaluate the ability of XLM-RoBERTa base, one of the leading Neural Language Models (NLMs), in distinguishing coherent text from incoherent text, where the latter has been artificially created to gradually undermine local coherence within text. Our findings indicate that NLMs still face challenges in modeling discourse coherence, and the linguistic knowledge acquired during the pre-training phase provides limited assistance when coherence relies on information not directly related to the topic. As expected, the cross-domain experiments highlighted that the model performances degrade with respect to the in-domain classification scenario, particularly when tested on data with a less defined structure, such as TED talks. Interestingly, the generalisation ability of the multilingual model holds across different languages, showing competitive results with the monolingual ones.

Limitations

We recognize the following main limitations of the present study. Although the approach we devised is not bounded to a specific model architecture and language, our study fouced only on one neural language model and a limited set of languages and this may limit the generalization of our results. Moreover, we are aware that discourse coherence is a multifactorial phenomenon that can only be partially covered by the devised methodology and dataset.

Ethics Statement

Our work has limited ethical implications since we mainly introduced an approach to study discourse coherence in NLMs. The datasets we built were used in compliance with the Terms of Use and the resources and materials produced during this study will be open source.

References

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Regina Barzilay and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China. Association for Computational Linguistics.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Youmna Farag, Josef Valvoda, Helen Yannakoudakis, and Ted Briscoe. 2020. Analyzing neural discourse coherence models. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 102–112, Online. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local

coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González, Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. 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. *CoRR*, abs/1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4328–4339.

Panitan Muangkammuen, Sheng Xu, Fumiyo Fukumoto, Kanda Runapongsa Saikaew, and Jiyi Li. 2020. A neural local coherence analysis model for clarity text scoring. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2138–2143, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proc. Interspeech 2021*, pages 3655–3659.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640.

Teun Adrianus Van Dijk. 1977. *Text and Context: Exploration in the Semantics and Pragmatics of Discourse*. Longman, London.

A Data Sample

ID	Source	Prompt	Target	Position	Class
66#193	TED	I think to be truly emotionally intelligent, we need to understand where those words have come from, and what ideas about how we ought to live and behave they are smuggling along with them. Let me tell you a story. It begins in a garret in the late 17th century, in the Swiss university town of Basel.	Inside, there's a dedicated student living some 60 miles away from home.	cont.	1
1158#40	TED	Unfortunately, global carbon emissions from deforestation now equals the global transportation sector. That's all ships, airplanes, trains and automobiles combined. So it's understandable that policy negotiators have been working hard to reduce deforestation, but they're doing it on landscapes that are hardly known to science.	It's like cutting a cake except this cake is about whale deep.	+5	0
591#383	TED	Or they mention cube roots or ask me to recite a long number or long text. I hope you'll forgive me if I don't perform a kind of one-man savant show for you today. I'm going to talk instead about something far more interesting than dates of birth or cube roots – a little deeper and a lot closer, to my mind, than work.	I'm asking you to do this because I believe our personal perceptions, you see, are at the heart of how we acquire knowledge.	+10	0
622#53	TED	But the really cool thing about them is when we put them together. You see that really salty Play-Doh? Well, it conducts electricity.	One of the most remarkable things about "Crime and Punishment" is its ability to thrill despite the details of the central murder being revealed in the first act.	out	0
1158#40	TED	Unfortunately, global carbon emissions from deforestation now equals the global transportation sector. That's all ships, airplanes, trains and automobiles combined. So it's understandable that policy negotiators have been working hard to reduce deforestation, but they're doing it on landscapes that are hardly known to science.	So our imagery is 3D, it's chemical, it's biological, and this tells us not only the species that are living in the canopy, but it tells us a lot of information about the rest of the species that occupy the rainforest.	-5	0
1887#464	TED	Now, you can think of that as the backbone that holds the rest of the molecule together. The three long chains on the right are called fatty acids, and it's subtle differences in the structures of these chains that determine whether a fat is, let's say, solid or liquid; whether or not it goes rancid quickly; and, most importantly, how good or how bad it is for you. Let's take a look at some of these differences.	Thank you for having me.	-10	0

Table 4: TED Data sample from the forward configuration.

ID	Source	Prompt	Target	Position	Class
680#11#6	WIKI	Its hair is short on its head and tail; however its legs tend to have longer hair. The hair on the majority of its body is grouped in clusters of 3-4 hairs. The hair surrounding its nostrils is dense to help filter particulate matter out as it digs.	Its tail is very thick at the base and gradually tapers.	cont.	1
31655#67#3	WIKI	Forthcoming soldiers consisted primarily of draftees or paid substitutes as well as poor enlistees lured by enlistment bonuses. The officers, however, were of a higher quality, responding out of a sense of civic duty and patriotism, and generally critical of the rank and file. Most of the 13,000 soldiers lacked the required weaponry; the war department provided nearly two-thirds of them with guns.	Nevertheless, the militia continued to deteriorate and twenty years later, the militia's poor condition contributed to several losses in the War of 1812, including the sacking of Washington, D.C., and the burning of the White House in 1814.	+5	0
14021#55#3	WIKI	Before this, the Copernican model was just as unreliable as the Ptolemaic model. This improvement came because Kepler realized the orbits were not perfect circles, but ellipses. Galileo Galilei was among the first to use a telescope to observe the sky, and after constructing a 20x refractor telescope.	While he was able to avoid punishment for a little while he was eventually tried and pled guilty to heresy in 1633.	+10	0
37914#20#3	WIKI	The libretto was prepared in accordance with the conventions of "opéra comique", with dialogue separating musical numbers. It deviates from Mérimée's novella in a number of significant respects. In the original, events are spread over a much longer period of time, and much of the main story is narrated by José from his prison cell, as he awaits execution for Carmen's murder.	In addition to DCI's work, the National Association of Theatre Owners released its Digital Cinema System Requirements.	out	0
27633#29#3	WIKI	The most common theory of how prehistoric people moved megaliths has them creating a track of logs which the large stones were rolled along. Another megalith transport theory involves the use of a type of sleigh running on a track greased with animal fat. Such an experiment with a sleigh carrying a 40-ton slab of stone was successfully conducted near Stonehenge in 1995.	The excavated remains of culled animal bones suggest that people may have gathered at the site for the winter rather than the summer.	-5	0
4183#39#3	WIKI	Products made from cellulose include rayon and cellophane, wallpaper paste, biobutanol and gun cotton. Sugarcane, rapeseed and soy are some of the plants with a highly fermentable sugar or oil content that are used as sources of biofuels, important alternatives to fossil fuels, such as biodiesel. Sweetgrass was used by Native Americans to ward off bugs like mosquitoes.	Others are simple derivatives of botanical natural products.	-10	0

Table 5: Wikipedia Data sample from the forward configuration.

B Detailed In-domain Classification Scores

	EN			ES			FR			IT			PT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
FW-10	0.71	0.71	0.70	0.72	0.72	0.72	0.71	0.71	0.71	0.71	0.71	0.71	0.73	0.73	0.73
	0.52	*0.52*	*0.52*	*0.52*	*0.52*	*0.51*	*0.54*	*0.54*	*0.54*	*0.52*	*0.52*	*0.52*	*0.55*	*0.55*	*0.55*
FW-5	0.72	0.71	0.71	0.69	0.69	0.69	0.67	0.67	0.66	0.64	0.64	0.64	0.68	0.68	0.68
	0.5	*0.5*	*0.49*	*0.53*	*0.52*	*0.52*	*0.53*	*0.53*	*0.53*	*0.54*	*0.54*	*0.54*	*0.54*	*0.54*	*0.54*
FW+5	0.67	0.67	0.66	0.67	0.67	0.67	0.64	0.64	0.64	0.67	0.67	0.67	0.67	0.67	0.67
	0.53	*0.53*	*0.53*	*0.52*	*0.52*	*0.52*	*0.51*	*0.51*	*0.51*	*0.53*	*0.53*	*0.53*	*0.55*	*0.55*	*0.55*
FW+10	0.72	0.72	0.72	0.73	0.72	0.72	0.73	0.73	0.73	0.68	0.68	0.68	0.73	0.73	0.73
	0.54	*0.54*	*0.53*	*0.51*	*0.51*	*0.51*	*0.52*	*0.52*	*0.52*	*0.52*	*0.52*	*0.52*	*0.55*	*0.55*	*0.55*
FW_OUT	0.96	0.96	0.96	0.95	0.95	0.95	0.94	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
	0.53	*0.53*	*0.52*	*0.51*	*0.51*	*0.51*	*0.52*	*0.52*	*0.52*	*0.53*	*0.53*	*0.53*	*0.52*	*0.52*	*0.52*
BW-10	0.71	0.71	0.71	0.71	0.71	0.7	0.69	0.69	0.69	0.70	0.70	0.70	0.71	0.71	0.71
	0.51	*0.51*	*0.51*	*0.52*	*0.52*	*0.52*	*0.55*	*0.55*	*0.54*	*0.53*	*0.53*	*0.53*	*0.53*	*0.53*	*0.53*
BW-5	0.69	0.69	0.69	0.68	0.68	0.68	0.64	0.64	0.64	0.64	0.64	0.64	0.66	0.66	0.66
	0.5	*0.5*	*0.5*	*0.51*	*0.51*	*0.51*	*0.54*	*0.54*	*0.54*	*0.54*	*0.54*	*0.54*	*0.56*	*0.56*	*0.56*
BW+5	0.67	0.67	0.67	0.68	0.68	0.68	0.67	0.67	0.67	0.66	0.66	0.66	0.64	0.63	0.63
	0.52	*0.52*	*0.52*	*0.52*	*0.52*	*0.51*	*0.54*	*0.54*	*0.53*	*0.54*	*0.54*	*0.53*	*0.54*	*0.54*	*0.54*
BW+10	0.74	0.73	0.73	0.70	0.70	0.70	0.72	0.71	0.71	0.71	0.7	0.70	0.72	0.71	0.71
	0.53	*0.53*	*0.53*	*0.52*	*0.52*	*0.52*	*0.54*	*0.54*	*0.53*	*0.53*	*0.53*	*0.53*	*0.54*	*0.54*	*0.54*
BW_OUT	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
	0.53	*0.53*	*0.53*	*0.5*	*0.5*	*0.5*	*0.53*	*0.53*	*0.53*	*0.54*	*0.54*	*0.54*	*0.53*	*0.53*	*0.53*

Table 6: Detailed in-domain classification scores reported by the xlm-RoBERTa-base model on Wikipedia data. Baseline scores are in italic.

	EN			ES			FR			IT			PT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
FW-10	0.69	0.69	0.69	0.7	0.7	0.69	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
	0.49	*0.49*	*0.48*	*0.51*	*0.51*	*0.51*	*0.49*	*0.5*	*0.37*	*0.47*	*0.48*	*0.44*	*0.5*	*0.5*	*0.49*
FW-5	0.69	0.69	0.69	0.66	0.66	0.66	0.61	0.61	0.61	0.67	0.67	0.66	0.66	0.66	0.66
	0.48	*0.49*	*0.43*	*0.63*	*0.5*	*0.34*	*0.5*	*0.5*	*0.34*	*0.56*	*0.51*	*0.39*	*0.5*	*0.5*	*0.36*
FW+5	0.66	0.66	0.66	0.63	0.63	0.63	0.61	0.61	0.61	0.67	0.67	0.66	0.64	0.64	0.64
	0.48	*0.48*	*0.46*	*0.57*	*0.51*	*0.35*	*0.54*	*0.5*	*0.34*	*0.56*	*0.51*	*0.38*	*0.44*	*0.49*	*0.35*
FW+10	0.69	0.69	0.69	0.67	0.67	0.67	0.66	0.66	0.66	0.68	0.68	0.68	0.68	0.68	0.68
	0.48	*0.49*	*0.43*	*0.53*	*0.51*	*0.39*	*0.48*	*0.5*	*0.36*	*0.53*	*0.52*	*0.45*	*0.4*	*0.49*	*0.34*
FW_OUT	0.84	0.83	0.83	0.84	0.83	0.83	0.82	0.82	0.82	0.83	0.83	0.83	0.83	0.83	0.83
	0.47	*0.48*	*0.45*	*0.51*	*0.51*	*0.47*	*0.51*	*0.5*	*0.41*	*0.5*	*0.5*	*0.49*	*0.43*	*0.49*	*0.35*
BW-10	0.67	0.67	0.67	0.65	0.65	0.65	0.66	0.65	0.64	0.64	0.64	0.64	0.67	0.67	0.67
	0.47	*0.48*	*0.43*	*0.59*	*0.51*	*0.38*	*0.25*	*0.5*	*0.33*	*0.55*	*0.51*	*0.39*	*0.31*	*0.49*	*0.33*
BW-5	0.62	0.62	0.62	0.61	0.6	0.59	0.62	0.61	0.61	0.63	0.63	0.63	0.66	0.66	0.65
	0.47	*0.49*	*0.4*	*0.75*	*0.5*	*0.34*	*0.5*	*0.5*	*0.34*	*0.64*	*0.51*	*0.35*	*0.5*	*0.5*	*0.35*
BW+5	0.65	0.65	0.65	0.63	0.62	0.62	0.65	0.64	0.63	0.63	0.63	0.63	0.66	0.66	0.66
	0.48	*0.49*	*0.45*	*0.6*	*0.51*	*0.36*	*0.25*	*0.5*	*0.33*	*0.54*	*0.5*	*0.35*	*0.25*	*0.49*	*0.33*
BW+10	0.68	0.68	0.68	0.64	0.64	0.63	0.68	0.67	0.67	0.68	0.68	0.67	0.64	0.64	0.64
	0.5	*0.5*	*0.41*	*0.55*	*0.51*	*0.41*	*0.32*	*0.49*	*0.33*	*0.52*	*0.51*	*0.42*	*0.36*	*0.49*	*0.33*
BW_OUT	0.82	0.82	0.82	0.82	0.81	0.81	0.83	0.83	0.83	0.83	0.83	0.83	0.85	0.85	0.85
	0.5	*0.5*	*0.47*	*0.52*	*0.52*	*0.52*	*0.49*	*0.5*	*0.37*	*0.51*	*0.51*	*0.49*	*0.5*	*0.5*	*0.34*

Table 7: Detailed in-domain classification scores reported by the xlm-RoBERTa-base model on TED data. Baseline scores are in italic.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The section "Limitations"

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The section "Abstract" and section 1 ("Introduction")

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We created two datasets and fine-tuned NLMs

☑ B1. Did you cite the creators of artifacts you used?
Sections: 1 and 2.1 and 2.3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section "Ethic Statement"

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
we used the datasets in compliance with the Terms of Use and the resources and materials produced during this study will be open source

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We cited the official references of the data and models used

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We reported such information in the section 2.1

C ☑ Did you run computational experiments?

2.2 "Experimental settings"

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
2.2 "Experimental settings"

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2.2 "Experimental settings"

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3 "Results" and "Appendix"

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We used default parameters

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.