Coherence Evaluation in Italian Language Models

Marta Sartor[1], Felice Dell'Orletta[1] and Giulia Venturi[1]

[1]Istituto di Linguistica Computazionale "A. Zampolli", (ILC-CNR) ItaliaNLP Lab, via G. Moruzzi 1, Pisa, Italy

Abstract

Coherence assessment is central to many NLP tasks, but its evaluation is complex and often done indirectly. In the LLM era, it is even more crucial to understand how, and how well, these models represent coherence. This study investigates the effectiveness of small Italian language models (under 1B parameters) in assessing coherence and focuses on what factors most influence their performance. Our analysis involves 15 Transformer-based LLMs differing in architecture, parameter size, and training data, and monitors different textual genres and perturbations used during dataset construction. Two coherence modeling strategies are tested: perplexity and inter-sentence semantic distance. We show that best practices vary significantly depending on model architecture and approach, but most importantly on what kind of texts they are applied to, highlighting the nuanced interaction between textual genre, data perturbation, and model performance.

Keywords

Italian LM, coherence assessment, perplexity, inter-sentence semantic distance

1. Introduction

Coherence is the meaning connection that binds the components of a text [2] and is fundamental to ensuring the effectiveness of every communicative act. Consequently, in computational linguistics, its analysis is crucial for the resolution of numerous tasks, from identifying the necessary information for question answering [3] to recognizing pathological speech [4, 5], from automatic readability assessment [6] to automatic summary generation [7]. Its critical importance has led to the development of a number of resources (e.g. [8, 9, 10]); for Italian specifically, a new dataset annotated with human judgments of coherence has recently been released (DisCoTex, [11]).

It is however notably complex to model coherence computationally, as it does not require explicit linguistic structures to be expressed: it is rather a psychological construct [12], reconstructed implicitly through inferences, general knowledge, co-text, and context [2]. Moreover, its highly subjective nature [11] makes coherence also difficult to assess and evaluate: the soundest approach, and the most direct, would be employing human evaluations, but such data is very costly and lengthy to collect. For this reason, the most common coherence evaluation strategies are by proxy, primarily through the order discrimination task. Its underlying assumption that shuffled texts are less coherent than the original, though sound [13], has shown its limits [14, 12, 15, 16] from the onset. However, its efficiency in terms of resources has encouraged several variations on the original task [17], ranging from altering the number and position of the shuffled sentences [15, 18] to replacing shuffling with substitutions from a closely related document [10].

Since the introduction of Transformer models and the paradigm shift brought about by Large Language Models, coherence modeling approaches have changed and shifted in that direction. Many works employ LMs, developing specific models through specialized training [19, 20] or leveraging pretrained models through new indirect approaches [21, 22]. A great deal of attention has since also been devoted to probing these models to more accurately evaluate their ability on coherence assessment [10, 23, 18].

Our contribution. We test the ability of small (under 1B parameters) Italian LLMs to model coherence, evaluating them against human judgments on two modeling strategies: perplexity and inter-sentence semantic distance. We were interested in examining how heavily different factors impact model performance, and we directed our analysis on three main aspects:

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]

marta.sartor@ilc.cnr.it (M. Sartor); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); giulia.venturi@ilc.cnr.it (G. Venturi)

- LLM characteristics;
- textual genre of the target text;
- textual perturbation applied during dataset construction.

The first point has a widely known impact and we attempted, compatibly with available models, to systematically monitor several components. To this end we selected 15 Transformer-based LLMs, all under 1 billion parameters, differing for architecture, parameter size, target language, and/or training data size. The literature is also quite clear on the impact that both textual genre and data perturbation can have, both in the training and the evaluation phase. Nonetheless, it is not always easy to monitor for these factors, especially due to resource availability. In order to take a deeper look at both these aspects, we chose to work on the DisCoTex [11] dataset, which contains small paragraphs from two different genres (TEDx and Wikipedia) and with different degrees of perturbation at the intersentential level (none, inversion, substitution), where each instance is annotated with human judgments of coherence.

2. Methodology

We tested two different unsupervised approaches to modeling coherence: *inter-sentence semantic distance* and *perplexity*.

The first approach is a widely used technique to compare vectorial representations of meaning and was calculated on all models tested, regardless of architecture. The paragraph is first divided into sentences using the sentence-splitting feature of the Stanza tokenizer, version 1.5.0[1]. Each sentence is tokenized and processed by the model, from which a single vector representation of the sentence is obtained: inter-sentence distance is then calculated between all pairs of consecutive sentences and a global paragraph value is obtained through a statistical function. Sentence embeddings were calculated differently on the basis of the model's architecture: for sentence encoders, the direct output of the model was taken; for decoders, the last layer representations of each token in the sentence were mean-pooled into a single vector; for encoder-decoders and encoders, the same process was applied to the encoder's last layer, and CLS was also tested as a possible sentence embedding for encoder models.
Being a less straightforward methodology, at each step we tested several variants to broaden the analysis as much as possible:

- the measure of distance was calculated both with cosine and Euclidean distance;
- for encoders, sentence embeddings were represented both through the CLS token and the mean-pooling of the sentence tokens;
- paragraph values were pooled from inter-sentence values by different statistical functions, namely mean and standard deviation.

Our choice of statistical function fell firstly of the mean, as a global measure of semantic distance widely used and recognized in the literature. We chose to add standard deviation, a less commonly used function, due to the nature of the dataset, where the data is locally perturbed: we felt that perhaps a local measure of distance could be suited to model such an alteration.

We also tried modeling coherence in terms of textual plausibility by using *perplexity*: it is a global indicator and has already been successfully employed to this end [12, 18]. This method was applied on all decoders and also on two selected encoder models which allowed a direct comparison with respect to the target language strategy. On decoder models we calculated perplexity, while with encoder models we used a plausibility metric so as to be able to compare it with perplexity. Plausibility was calculated through masked language modeling by masking each token of the paragraph one by one and averaging their likelihood across the paragraph, as reported below with n being the total number of tokens:

$$plausibility(X) = \frac{1}{n}\sum_{i=1}^{n} P(x_i)$$

[1] https://pypi.org/project/stanza/1.5.0/#files

In order to address the impact of textual genre and text perturbation, the analysis of the results was carried out at various levels of granularity: on the entire dataset, by source, and by perturbation for each source. Each model was evaluated based on the Spearman correlation of its results with human judgment. Additionally, the difference in distribution between classes (source of texts or type of perturbation) was assessed using the Wilcoxon T-test and rank biserial correlation. Evaluating performance through correlation with human judgment, besides being a more straightforward and reliable approach, also allows us to effectively counterbalance the possible bias introduced by the fact that Wikipedia is present both in the pretraining of most models and in our evaluation dataset.

Baselines were set as random values. Perplexity and plausibility were both assimilated to a probability distribution and thus we generated random values between 0 and 1. For inter-sentence distance, the chosen measure of distance was calculated between as many random values as the average length in sentences of the dataset items, which is 4; the range in which we generated each value changed based on the measure: -1 to 1 for cosine, and 0 to 1 for Euclidean distance (as if the values had been normalized, since it has an infinite maximum).

2.1. Dataset

In this work, we used the dataset released by [11], recently integrated into a larger benchmark released for DiSCoTex [24], a shared task on textual coherence analysis task presented at the 8th evaluation campaign of NLP and speech tools for the Italian language (EVALITA 2023) [25]. This dataset consists of 1064 instances, each corresponding to a paragraph of 4-5 sentences and annotated with human coherence judgments. The data is sourced either from the Italian Wikipedia or the Italian section of the Multilingual TEDx dataset (which contains TEDx transcripts), to represent different linguistic varieties; the instances are balanced for source.

During the dataset construction about 2/3rds of the instances were subjected to alterations that more or less significantly damaged the internal coherence of the paragraph, to test the effect on human judgment of some common text perturbation strategies. The alterations were either the inversion of any two sentences within the paragraph (inversion perturbation), or the replacement of a sentence in the paragraph with the tenth sentence from the end of the paragraph (substitution perturbation). The remaining third of the instances was instead left unaltered, to serve as a control group.

Each paragraph is annotated with human judgment values, corresponding to the mean and standard deviation of the ratings collected on each instance from at least 10 human annotators. The judgments were collected through crowdsourcing from native Italian speakers and are expressed on a Likert scale from 1 to 5, 1 being the lowest coherence score and 5 the highest.
It must be noted, however, that the source or perturbation type differentiates texts significantly on the basis of the distribution of human coherence judgment they receive (see figures 1 and 2), to the point that the difference between distributions remains statistically significant even when differentiating instances for both source and perturbation type (see figure 3). Indeed, the difference increases the heavier the perturbation applied on the instance, but the source has a much stronger impact than perturbation. It is also worth noting that the value range in which most Wikipedia texts are located is far less sparse than that occupied by texts sourced from TEDx.

For this work, the dataset was integrated with additional data which were not available in the released version, namely all source and perturbation labels.

2.2. Models

We tested 15 different Transformer-based models, covering the most common architectures (encoder, decoder, encoder-decoder, sentence encoder).

Most models are BERT-based to allow better comparability on some of the variations we wanted to account for: we used a multilingual version (mBERT) and two monolingual Italian versions (BERT-ita and BERT-ita-xxl) differentiated by dataset size, as well as a sentence encoder (sBERT). Different architecture sizes were not tested because they were not available for the Italian version of these models;

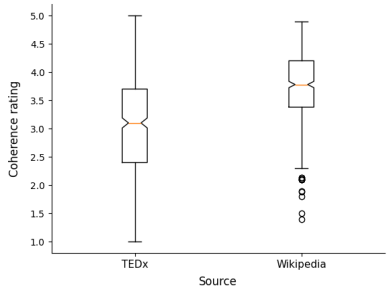

Figure 1: Mean coherence value distribution on the basis of textual genre.

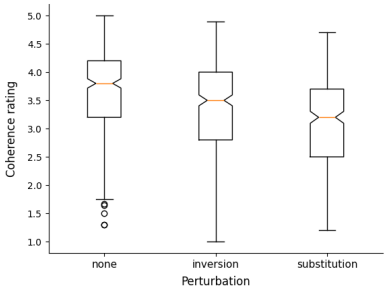

Figure 2: Mean coherence value distribution on the basis of perturbation type.

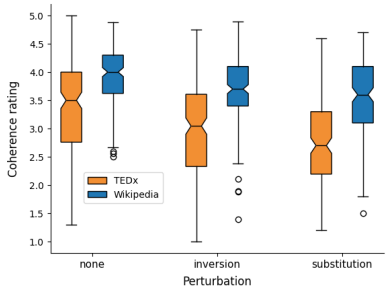

Figure 3: Mean coherence value distribution on the basis of genre and perturbation type.

Table 1
Descriptive table of the models tested, highlighting the dimensions that we aim to compare. n.d. means that the training data size was not declared.

	architecture	parameter size	training data (size)	language
LABSE [26]	sentence encoder	470M	n.d.	multilingual
MUSE [27]	sentence encoder	69M	n.d.	multilingual
MUSE large[2]	sentence encoder	85M	n.d.	multilingual
sBERT[3]	sentence encoder	111M	n.d.	italian
mBERT[4]	encoder	179M	n.d.	multilingual
BERT-ita[5]	encoder	111M	13GB	italian
BERT-ita-xxl[6]	encoder	111M	81GB	italian
XLM-R base [28]	encoder	250M	2.5T	multilingual
XLM-R large [28]	encoder	560M	2.5T	multilingual
IT5 small [29]	encoder-decoder	60.5M	215GB	italian
IT5 base [29]	encoder-decoder	223M	215GB	italian
IT5 large [29]	encoder-decoder	770M	215GB	italian
GroGPT [30]	decoder	117M	13.8GB	italian
GePpeTto [31]	decoder	117M	13.8GB	italian
Minerva[7]	decoder	350M	n.d.	italian

uncased versions were not tested due to the nature of the Italian language and, most importantly, the developers' own recommendations which indicated that the cased version was better.

Table 1 summarizes the models' characteristics with respect to our research questions. With "language" here we refer to the target language of the models, which does not always coincide with the language of training data: for example, GroGPT is developed for Italian but is an English GPT-2 model whose lexical embeddings have been retrained for Italian.

3. Experimental Results

As previously stated, the coherence judgments expressed by people are on a Likert scale from 1 (not very coherent) to 5 (very coherent). Cosine and Euclidean distances, as well as perplexity, conversely express greater coherence when the score is lower, so a negative correlation is expected. Plausibility, on the other hand, has low scores for incoherent texts and high scores for coherent texts like the Likert scale, being a probability distribution; thus, the direction of the correlation is opposite to those of all other

[2]https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
[3]https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased
[4]https://github.com/google-research/bert/blob/master/multilingual.md
[5]https://huggingface.co/dbmdz/bert-base-italian-cased
[6]https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[7]https://huggingface.co/sapienzanlp/Minerva-350M-base-v1.0

Table 2
Spearman correlation of human judgment labels with model predictions, on the entire dataset. Except for sentence encoders, the sentence pooling strategy is mean-pooling unless stated otherwise. The asterisk indicates p-value < 0,05.

	mean eucl ↓	mean cos ↓	std eucl ↓	std cos
baseline	0,01	-0,07	0,03	-0,03
LABSE	***-0,43**	***-0,43**	*0,06	0,01
MUSE	*-0,37	*-0,38	*0,09	0,05
MUSE large	*-0,38	*-0,39	*0,07	0,03
sBERT	***-0,42**	***-0,45**	0,01	-0,02
mBERT CLS	*-0,14	*-0,14	*-0,06	*-0,09
mBERT	*-0,29	*-0,34	-0,04	*-0,10
BERT-ita CLS	-0,02	*-0,14	-0,03	*-0,12
BERT-ita-xxl CLS	*-0,20	*-0,24	*-0,09	*-0,15
BERT-ita	*-0,28	*-0,32	0,03	-0,05
BERT-ita-xxl	*-0,38	*-0,39	*-0,13	*-0,13
XLM-R base	*-0,32	*-0,32	*-0,19	*-0,23
XLM-R large	*-0,34	*-0,33	*-0,18	*-0,22
IT5 small	*-0,38	*-0,37	*-0,17	*-0,19
IT5 base	*-0,36	*-0,36	*-0,20	*-0,23
IT5 large	*-0,36	*-0,37	*-0,19	*-0,14
GroGPT	*-0,13	*-0,18	*-0,10	*-0,13
GePpeTto	*-0.20	*-0.34	-0.05	*-0.10
Minerva	*-0,35	0,00	*-0,28	*-0,10

measures. In order to compare plausibility and perplexity, the sign of the correlation with plausibility has been inverted; we will henceforth refer to it as pseudo-perplexity.

The analysis of results was performed on three different levels: on the entire dataset (sect. 3.1) and, to account for genre and perturbation differences, separating instances by their source (sect. 3.2) and for each source by their perturbation type (sect. 3.3).

3.1. Overall Analysis

The Spearman correlation of the models' prediction and human judgments of coherence is shown in tables 2 and 3, which cover approaches using inter-sentence distance and (pseudo)perplexity respectively. Coefficients marked with an asterisk are statistically significant.
Results for each tested methodology and model are always above the baseline, except for the inter-sentence standard deviation of some models. Globally, the strongest correlation with human judgment is obtained by perplexity with Minerva (-0.46) and mBERT (-0.45), and by the average inter-sentence cosine distances calculated with sBERT (-0.45). Among the best are also the average cosine (-0.43) and Euclidean (-0.43) distances with LaBSE, and the average Euclidean distance with sBERT (-0.42).

Comparing the different functions of inter-sentence distance, standard deviation appears to be an unreliable approach: results often lacked correlation with human judgment or had very low coefficients compared to mean inter-sentence distance, which is also (almost) always statistically significant. Across all models, the standard deviation averages -0.08 and -0.11 correlation for Euclidean and cosine distance respectively, while the mean inter-sentence distance averages, respectively, -0.30 and 0.31. These results also highlight a different trend, which is that Euclidean distance generally obtains lower scores than cosine distance. This difference is even more pronounced (-0.29 vs -0.32, -0.07 vs -0.11) when leaving out the one notable exception, Minerva, which despite great results with Euclidean distances has 0 to non-significant correlation with cosine distance. The best approach however remains using perplexity or pseudo-perplexity (-0.38), whose average is much higher than what the same models averaged when using mean inter-sentence distance (Euclidean: -0.27, cosine: -0.25). For what concerns sentence encodings, embeddings obtained with CLS are significantly worse, not only with respect to

the corresponding mean-pooled embeddings (as already suggested by the literature: see e.g. [32]) but also to every other model.

Table 3
Spearman correlation between (pseudo)perplexity and human judgment labels, on the entire dataset. The asterisk indicates p-value < 0,05.

	(P)PPL ↓
baseline	-0,06
mBERT	***-0,45**
BERT-ita-xxl	*-0,35
GroGPT	*-0,39
GePpeTto	*-0,27
Minerva	***-0,46**

Among the different architectures, sentence encoders obtain overall the best results: they obtain the highest correlation scores and there is a great consistency among different models: their average mean inter-sentence distance coefficient is -0.40 for Euclidean distance and -0.41 for cosine distance, both higher than the average perplexity score. Their predictions have consistently lower correlation than other models (or no correlation at all) only when observing standard deviation of inter-sentence distance. On the other hand IT5, which represents encoder-decoders, has a good correlation with human judgment both by mean and standard deviation of inter-sentence distance, the latter being statistically significant and the highest among all models but rather low. The different IT5 versions have, with mean inter-sentence distance, lower correlation than sentence encoders, but still perform on par with the best encoder models. Overall, the correlation with human judgment of encoder results appears very variable. However, the high variability is due to the much lower performance obtained when using CLS, instead of mean-pooling, as a sentence representation method: this strategy results in little to no significant correlation with human judgments. Considering only encoders with the mean-pooling strategy, the average correlation with human judgment is fairly high, although inferior to that of sentence encoders and IT5. Decoders have the highest variability among results depending on the specific model: Minerva obtains the best results, but only using Euclidean distance, while methods employing cosine distance lead to non-significant results; GePpeTto has the lowest perplexity scores but performs well with mean inter-sentence distance; lastly, GroGPT has a low but consistent performance with both mean and standard deviation inter-sentence distance but has good perplexity scores.

For what concerns parameter size, this does not seem to influence results neither comparing different sizes of the same model (for those where the comparison is possible) nor considering the absolute parameter size: only with sentence encoder the correlation increases in parallel with parameter size, but the difference is rather small. Training dataset size, on the other hand, consistently changes the performance from BERT-ita to BERT-ita-xxl, especially when using CLS.

It is instead unclear how, or if, target language influences performance: multilingual encoders perform in the middle between BERT-ita and BERT-ita-xxl when considering inter-sentence distances, hinting at a possible advantage, and with (pseudo)perplexity multilingual BERT has a correlation score which is almost on par with that of the much bigger Minerva. In sentence encoders, however, the opposite seems to be true: sBERT performs comparably to the best multilingual sentence encoder, LABSE, despite considerable parameter size difference.

3.2. Analysis by source

Tables 4 and 5 show the correlation coefficients of model predictions with human judgments when dividing the dataset by source. Confirming the results of [24], performance on the TEDx and the Wikipedia sections is very different, with the first obtaining higher coefficients with all coherence assessment approaches (with the only exception of the mean inter-sentence cosine distance calculated with IT5 base). On Wikipedia, the correlation with human judgments is more likely to be not significant

and its coefficient is stronger than -0.2 only with sentence encoders or through (pseudo)perplexity scores: excluding standard deviation of inter-sentence distance, the average Wikipedia coefficient is -0.12, while for TEDx it is -0.24. This could be influenced by the fact that human coherence judgments on Wikipedia texts are more densely distributed in the upper (high-coherence) range, as exemplified in figure 1; it is also worth noting that higher coherence values for Wikipedia texts were also produced by almost all models regardless of approach, although the entity of this difference varied consistently between models.
In line with what was observed on the entire dataset, sentence encoders remain the best performing architecture and (pseudo)perplexity the most effective coherence assessment approach. The unsuitability of standard deviation of inter-sentence distance and using CLS as sentence encoder is also further confirmed, with cosine distance still obtaining better results than Euclidean distance. Minerva also maintains the skewness in results between approaches using Euclidean and cosine distances.

The combination that achieves the highest correlation with human judgment is perplexity calculated with Minerva (-0.46 and -0.26 on TEDx and Wikipedia respectively), followed closely by pseudo-perplexity calculated with BERT-ita-xxl (-0.43, -0.23); the average inter-sentence cosine distance calculated with sBERT (-0.38, -0.24) also obtains satisfying results.

As we already observed, standard deviation of inter-sentence distance is not a reliable coherence indicator: correlation with human judgment is hardly ever significant, and when it is, it is only significant on one of the two classes and with very low coefficients. Perplexity and pseudo-perplexity on the other hand, with the sole exception of GePpeTto, obtain much higher correlation on TEDx than any other approach and keep consistently high coefficients on Wikipedia, where most other performances falter. Besides GePpeTto, perplexity results vary significantly between models on TEDx (from 0.32 to 0.46) but are mostly identical (from -0.23 to -0.26) on Wikipedia; moreover, the best perplexity results gain a noticeable margin from those of inter-sentence distances (0.08) on TEDx, while those of the Wikipedia improve only 0.02. This reduced improvement brought about by perplexity, together with the overall lower performance on the Wikipedia section, supports our claim that its presence in most training datasets is offset by using human coherence judgments, and not perturbation labels, for the evaluation.

Sentence encoders remain the best performing architecture, not only for their higher correlation coefficients on the TEDx section but also and especially for their performance on the Wikipedia section, which is always significant and higher on average than any other architecture. As was the case on the overall dataset, their average score (-0.35 for TEDx and -0.19 for Wikipedia) is close to the average perplexity (-0.36 and -0.21 respectively), although this time slightly lower. Sentence encoders also appear in general much less sensitive than other architectures to the type of distance used, except for sBERT. Encoders maintain a certain variability depending on the models and are still comparable to decoders when using perplexity, but this time with inter-sentence distances they perform generally better than IT5, for which the Wikipedia class has always low to no correlation with human judgment. Decoders, on the other hand, have consistently low performances on inter-sentence distance tasks, in contrast with what previously observed; the sole exception is GePpeTto, when leveraging the mean cosine distance: once more his leading role as an encoder reverses as a decoder, where he obtains the lowest scores.

For what concerns parameters and training data size, no significant differences were observed. The impact of the target language is, however, unclear: overall there seems to be no clear preference for either, and the direct comparison between mBERT and BERT-ita/BERT-ita-xxl shows the former outperforming the latter on inter-sentence distance tasks and the opposite on pseudo-perplexity.

3.3. Analysis by source and perturbation

As we already observed, the huge differences between TEDx and Wikipedia (both in terms of human judgment and model behavior) are such that the impact of different kinds of perturbation can only be observed by keeping the two genres separate. The impact of the aforementioned difference can be very clearly seen also at this level of analysis: correlation results exhibit strong differences between the TEDx and Wikipedia sections, both in terms of significance and class distribution. Not only do the TEDx

Table 4
Spearman correlation of human judgment labels with model predictions, dividing the dataset by source. Except for sentence encoders, the sentence pooling strategy is mean-pooling unless stated otherwise. The asterisk indicates p-value < 0,05.

	MEAN ↓		STD ↓	
	TED	**WIKI**	**TED**	**WIKI**
LABSE eucl	*-0,36	*-0,21	0,05	0,07
LABSE cos	*-0,37	*-0,20	0,02	0,04
MUSE eucl	*-0,33	*-0,18	*0,14	0,05
MUSE cos	*-0,33	*-0,17	*0,12	0,02
MUSE large eucl	*-0,37	*-0,18	*0,14	0,04
MUSE large cos	*-0,37	*-0,18	*0,11	0,02
sBERT eucl	*-0,32	*-0,19	0,06	0,07
sBERT cos	***-0,38**	***-0,24**	0,03	0,04
mBERT CLS eucl	*-0,14	*-0,10	0,01	-0,03
mBERT CLS cos	*-0,13	*-0,12	-0,03	-0,06
mBERT eucl	*-0,24	*-0,10	-0,05	0,01
mBERT cos	*-0,28	*-0,16	-0,08	-0,03
BERT-ita CLS eucl	-0,06	-0,01	-0,06	-0,02
BERT-ita CLS cos	*-0,13	-0,06	*-0,11	-0,07
BERT-ita-xxl CLS eucl	*-0,09	-0,08	0,04	-0,05
BERT-ita-xxl CLS cos	*-0,11	-0,08	0,01	-0,08
BERT-ita eucl	*-0,17	*-0,09	0,02	0,05
BERT-ita cos	*-0,23	*-0,11	0,00	0,00
BERT-ita-xxl eucl	*-0,25	-0,08	-0,03	-0,04
BERT-ita-xxl cos	*-0,26	*-0,12	-0,04	-0,04
XLM-R base eucl	*-0,19	*-0,11	-0,08	-0,06
XLM-R base cos	*-0,18	*-0,12	*-0,10	-0,07
XLM-R large eucl	*-0,23	*-0,10	*-0,13	-0,02
XLM-R large cos	*-0,22	*-0,10	*-0,16	-0,04
IT5 small eucl	*-0,28	-0,07	-0,08	-0,06
IT5 small cos	*-0,25	*-0,09	-0,07	-0,06
IT5 base eucl	*-0,21	*-0,09	-0,08	-0,08
IT5 base cos	*-0,23	*-0,11	-0,07	*-0,12
IT5 large eucl	*-0,23	-0,06	*-0,09	-0,04
IT5 large cos	*-0,25	*-0,10	0,00	-0,07
GroGPT eucl	-0,07	-0,01	-0,07	0,02
GroGPT cos	-0,07	-0,05	-0,02	-0,03
GePpeTto eucl	*-0.20	-0.05	-0.08	0.00
GePpeTto cos	*-0.24	*-0.16	*-0.12	-0.03
Minerva eucl	*-0.20	-0.06	*-0.16	-0.06
Minerva cos	0.00	-0.03	-0.05	-0.04

section results show higher correlation coefficients, but in the Wikipedia section, most correlations are not significant. Furthermore, the perturbation class with the highest correlation with human judgment for TEDx, namely the inversion class, is never significant in the Wikipedia section except when using (pseudo)perplexity. It is worth noting that how the perturbation classes rank in terms of performance is not only different between TEDx and Wikipedia, and for Wikipedia between inter-sentence distance-based methods and perplexity, but also that these differences are not aligned with inter-annotator agreement. The only common factor between the two sections is the effectiveness of pseudo-perplexity and sentence encoders and the ineffectiveness of standard deviation of inter-sentence distance. Due to the significant differences, the two sections are treated separately.

Table 5
Spearman correlation between (pseudo)perplexity and human judgment labels, by source. The asterisk indicates p-value < 0,05

	(P)PPL ↓	
	TED	**WIKI**
mBERT	* -0,32	* -0,25
BERT-ita-xxl	* **-0,43**	* **-0,23**
GroGPT	* -0,34	* -0,24
GePpeTto	* -0,25	-0,09
Minerva	* **-0,46**	* **-0,26**

3.3.1. TEDx

The correlation of models' predictions and human judgment in the TEDx section is shown in tables 6 and 7, for inter-sentence distance approaches and (pseudo)perplexity respectively. Among the different perturbation classes, the inversion class generally has the strongest correlation with human judgment, generally followed by the class without alterations and then by the substitution class. Correlation is generally statistically significant except for standard deviation of inter-sentence distance measures, which as we already observed is an unreliable approach. The only other cases of non-significant coefficients are with mean inter-sentence distance, mainly in the substitution class and only in a few cases in the unaltered class, mostly when using CLS as sentence encoders.

The highest correlation with human judgment was obtained by Minerva's perplexity (-0.45 unaltered, -0.51 inversion, -0.39 substitution), followed by BERT-ita-xxl's pseudo-perplexity (-0.45, -0.45, -0.33) and LABSE's mean inter-sentence cosine distance (-0.35, -0.42, -0.25).

As always, perplexity was the approach with highest correlation to human judgment, although with considerable internal variability on the basis of the model: it averaged -0.35 for the unaltered class, -0.39 for the inversion class, and -0.30 for the substitution class. Also in line with previous observations, sentence encoders remain the best performing architecture, obtaining consistently high results and averaging -0.34, -0.36, and -0.23 for the unaltered, inversion, and substitution classes respectively, not too far from the perplexity scores. The role of the model language (multilingual or Italian) remains however unclear, following the same patterns observed in the dataset divided by source.

Generally speaking, this level of analysis is mostly coherent with the previous ones. Some differences concern the role of parameter and training size. While parameter size still does not seem relevant in absolute terms, this time it has a positive impact when considering different sizes of MUSE and XLM-R (although it seems almost counterproductive on IT5). Similarly, training data size improves performance from BERT-ita to BERT-ita-xxl (especially with mean-pooling), but does not seem to influence other models. The difference between Euclidean and cosine distances is also reduced.

3.3.2. Wikipedia

Tables 8 and 9 show the correlation between the model's predictions and the human coherence judgments. The most interesting results concern the perturbation classes: the performance ranking of the different classes is not only different from that of TEDx, but also different between approaches using inter-sentence distance and using (pseudo)perplexity. In the first case, the substitution class has the highest correlation with human judgment, while the inversion class performs the worst never being statistically significant; when using (pseudo)perplexity, on the other hand, the inversion class is the one which correlates the most with human judgment, followed by the substitution class. Upon further inspection, on the substitution and unaltered classes there is not much difference between the average performance using (pseudo)perplexity (-0.15 and -0.22 respectively) or mean inter-sentence distance (-0.14 and -0.17, excluding outliers like CLS embeddings and cosine Minerva), especially if considering cosine distance (-0.15 and -0.20). What changes, radically, is performance in the inversion class, going from no correlation to -0.25.

Table 6
Spearman correlation between human judgments and unsupervised methodologies tested on pretrained models. Results on the **TEDx** section of the dataset, grouped by perturbation type. Except for sentence encoders, the sentence pooling strategy is mean-pooling unless stated otherwise. The asterisk indicates p-value < 0,05.

	MEAN ↓			**STD ↓**		
	no	**swap**	**sub**	**no**	**swap**	**sub**
LABSE eucl	*-0,34	*-0,42	*-0,25	0,02	0,07	0,08
LABSE cos	***-0,35**	***-0,42**	***-0,25**	-0,03	0,03	0,05
MUSE eucl	*-0,31	*-0,35	*-0,20	0,07	*0,20	0,07
MUSE cos	*-0,32	*-0,35	*-0,20	0,04	*0,18	0,06
MUSE large eucl	*-0,39	*-0,34	*-0,25	0,09	0,13	0,13
MUSE large cos	*-0,40	*-0,34	*-0,26	0,06	0,11	0,12
sBERT eucl	*-0,30	*-0,30	*-0,20	-0,08	0,05	*0,20
sBERT cos	*-0,34	*-0,34	*-0,25	-0,06	-0,03	*0,20
mBERT CLS eucl	-0,07	-0,14	*-0,21	0,06	0,01	-0,11
mBERT CLS cos	-0,05	-0,14	*-0,20	0,05	-0,04	-0,14
mBERT eucl	*-0,23	*-0,20	*-0,25	-0,13	0,02	-0,08
mBERT cos	*-0,23	*-0,30	*-0,25	-0,10	-0,05	-0,09
BERT-ita CLS eucl	0,00	-0,08	-0,05	-0,05	0,02	*-0,22
BERT-ita CLS cos	-0,10	*-0,19	-0,06	-0,09	-0,09	*-0,18
BERT-ita-xxl CLS eucl	0,00	*-0,22	-0,04	-0,02	0,05	0,07
BERT-ita-xxl CLS cos	-0,01	*-0,24	-0,05	-0,05	0,00	0,05
BERT-ita eucl	*-0,19	*-0,16	*-0,16	-0,06	0,04	0,03
BERT-ita cos	*-0,24	*-0,25	*-0,17	-0,10	-0,01	0,03
BERT-ita-xxl eucl	*-0,26	*-0,32	-0,14	-0,12	-0,01	0,05
BERT-ita-xxl cos	*-0,24	*-0,34	-0,14	-0,09	-0,04	0,03
XLM-R base eucl	*-0,19	*-0,29	-0,05	*-0,17	-0,11	0,03
XLM-R base cos	*-0,19	*-0,27	-0,04	*-0,17	*-0,15	0,01
XLM-R large eucl	*-0,22	*-0,32	-0,13	*-0,25	*-0,16	-0,03
XLM-R large cos	*-0,23	*-0,31	-0,11	*-0,27	*-0,20	-0,05
IT5 small eucl	*-0,27	*-0,36	*-0,21	*-0,15	-0,06	-0,01
IT5 small cos	*-0,26	*-0,32	-0,15	*-0,17	-0,04	-0,01
IT5 base eucl	-0,14	*-0,38	-0,10	0,04	*-0,16	-0,15
IT5 base cos	*-0,21	*-0,40	-0,06	-0,02	-0,14	-0,09
IT5 large eucl	*-0,17	*-0,34	*-0,17	-0,03	*-0,18	-0,08
IT5 large cos	*-0,22	*-0,35	*-0,18	0,03	-0,07	0,00
GroGPT eucl	-0,02	*-0,18	-0,03	-0,13	-0,11	-0,02
GroGPT cos	-0,02	-0,12	-0,09	0,03	-0,09	-0,03
GePpeTto eucl	*-0,20	*-0,16	-0,10	-0,10	-0,07	0,02
GePpeTto cos	*-0,26	*-0,22	-0,12	-0,13	-0,11	-0,01
Minerva eucl	*-0,19	*-0,29	-0,10	*-0,17	*-0,22	-0,05
Minerva cos	0,01	0,03	0,05	-0,04	-0,14	0,11

There is an overall drop in performance, with an increased number of results that are not statistically significant. Comparing the different approaches, the pattern is the same as at the other levels of analysis: standard deviation is the worst, as it is almost never significant, and perplexity performs the best, especially since it is the only methodology where all three classes are statistically significant. Moreover, with (pseudo)perplexity all models (except for GePpeTto) always have statistically significant results, while with mean inter-sentence distance only about half of the results of the unaltered and the substitution class are statistically significant.

The highest correlation scores are obtained by Minerva with perplexity (-0.18 unaltered, -0.32 inversion, -0.26 substitution), mBERT with pseudo-perplexity (-0.21, -0.28 and -0.25 respectively), and sBERT with mean inter-sentence cosine distance (-0.28, -0.07, -0.34).

Results are in line with what we observed on the overall dataset and considering sources separately;

Table 7
Spearman correlation between (pseudo)perplexity and human judgment labels, by perturbation type on texts sourced from **TEDx**. The asterisk indicates coefficients with p-value < 0,05.

	(P)PPL ↓		
	no	**swap**	**sub**
mBERT	*-0,26	*-0,38	*-0,28
BERT-ita-xxl	***-0,45**	***-0,45**	***-0,33**
GroGPT	*-0,31	*-0,40	*-0,27
GePpeTto	*-0,26	*-0,23	*-0,22
Minerva	***-0,45**	***-0,51**	***-0,39**

sentence encoders, in particular, are the only models that manage to have two statistically significant classes with distance-based approaches. There is only a slight difference in what concerns the impact of language. When directly comparing mBERT and BERT-ita and BERT-ita-xxl, the first has better performances both with inter-sentence distance measures and pseudo-perplexity measures, and overall multilingual models seem to be performing better (except for sBERT among sentence encoders).

4. Conclusions

We evaluated the coherence assessment abilities of 15 small Italian language models, which varied in their structural and training-related characteristics, using two unsupervised approaches: modeling coherence based on inter-sentence semantic distance and perplexity. We evaluated results by their correlation with human judgment of coherence and analysed our dataset at different levels, to monitor differences related to the genre of the target text and the perturbation it was subjected to.

Perplexity and pseudo-perplexity consistently obtain the highest correlation with human judgments and seem to be the most effective coherence assessment methods. When considering distance measures, the accuracies obtained with sentence encoders were comparable to those of (pseudo)perplexity. Cosine distance appeared to be slightly better than Euclidean distance, while sentence embeddings through CLS and standard deviation of a paragraph's inter-sentence distance proved to be unsuitable. With perplexity and pseudo-perplexity the single most impactful decision seemed to be the model, regardless of parameter size or architecture; conversely, architecture was the most influential factor with inter-sentence distance approaches, with sentence encoders obtaining by far the best results. This was shown not only by higher correlation coefficients but also by the very close range of values produced by the models, underlining the reliability of the approach. Model and training set size did not seem to influence much performance, while model language (multilingual or Italian) had contradictory results.

Textual genre was shown to heavily influence model performance, both quantitatively and qualitatively, with TEDx always obtaining much higher correlation coefficients than Wikipedia. It is unlikely that they have been influenced by the presence of Wikipedia in the training, given both the lower results and the evaluation against human judgments. These results could instead be influenced by the wider value range in human judgments registered on the former, aiding a ranking-based correlation measure, which underlines the relevance of considering genre in performance evaluation.

The impact of different sources is also clear when the effect of perturbations is analyzed. Perturbation classes not only exhibited markedly different behavior but also had different results depending on the source of the paragraph. The clearest example is that of the inversion class, which performed the best on TEDx, while on Wikipedia obtained good results with (pseudo)perplexity but was never statistically significant with distance measures. Inversions impact order, which is more easily picked up by a sequence-based metric like perplexity than by a semantically-rooted distance measure. Both were enough to pick up alterations on TEDx, but only the former was effective on Wikipedia due to its higher thematic coherence, highlighting the importance of considering the perturbations both in isolation and in their interaction with other textual characteristics.

Table 8
Spearman correlation between human judgments and unsupervised methodologies tested on pretrained models. Results on the **Wikipedia** section of the dataset, grouped by perturbation type. Except for sentence encoders, the sentence pooling strategy is mean-pooling unless stated otherwise. The asterisk indicates p-value < 0,05.

	MEAN ↓			STD ↓		
	no	**swap**	**sub**	**no**	**swap**	**sub**
LABSE eucl	*-0,22	-0,08	*-0,29	0,11	0,01	0,05
LABSE cos	*-0,21	-0,08	*-0,29	0,08	0,01	0,03
MUSE eucl	*-0,15	-0,09	*-0,24	0,07	-0,05	0,11
MUSE cos	*-0,15	-0,09	*-0,24	0,04	-0,07	0,07
MUSE large eucl	*-0,15	-0,05	*-0,27	0,13	-0,04	0,01
MUSE large cos	-0,14	-0,05	*-0,27	0,12	-0,04	-0,01
sBERT eucl	***-0,28**	-0,06	***-0,26**	0,10	0,03	0,09
sBERT cos	*-0,28	-0,07	*-0,34	0,09	0,00	0,07
mBERT CLS eucl	-0,10	0,02	*-0,17	-0,11	0,03	-0,05
mBERT CLS cos	-0,12	0,00	*-0,19	-0,13	0,00	-0,07
mBERT eucl	-0,11	0,06	*-0,16	0,01	0,02	0,03
mBERT cos	*-0,16	-0,02	*-0,26	-0,05	-0,04	0,02
BERT-ita CLS eucl	-0,03	0,02	-0,01	-0,09	0,12	-0,07
BERT-ita CLS cos	-0,08	0,01	-0,10	-0,12	0,07	*-0,18
BERT-ita-xxl CLS eucl	*-0,20	0,00	-0,03	-0,12	-0,01	-0,05
BERT-ita-xxl CLS cos	*-0,21	0,00	-0,04	-0,14	-0,03	-0,10
BERT-ita eucl	-0,10	0,00	-0,09	0,01	0,07	0,11
BERT-ita cos	-0,11	0,01	*-0,16	0,03	0,04	0,02
BERT-ita-xxl eucl	-0,14	0,06	-0,13	-0,07	0,07	-0,03
BERT-ita-xxl cos	-0,13	0,02	*-0,20	-0,04	0,04	-0,04
XLM-R base eucl	*-0,16	0,01	*-0,17	-0,04	-0,03	-0,12
XLM-R base cos	*-0,16	0,01	*-0,16	-0,06	-0,03	-0,14
XLM-R large eucl	*-0,15	0,00	-0,10	-0,04	0,00	0,04
XLM-R large cos	*-0,16	0,00	-0,09	-0,06	0,01	0,01
IT5 small eucl	-0,12	0,00	-0,11	-0,06	0,04	-0,13
IT5 small cos	-0,11	-0,01	-0,15	-0,07	0,02	-0,09
IT5 base eucl	-0,14	-0,03	-0,14	-0,09	0,03	*-0,22
IT5 base cos	*-0,15	-0,05	*-0,16	-0,12	-0,03	*-0,23
IT5 large eucl	-0,06	0,03	*-0,16	-0,03	0,06	-0,13
IT5 large cos	-0,12	-0,01	*-0,17	-0,04	0,01	-0,09
GroGPT eucl	0,03	-0,10	0,09	0,04	0,01	0,05
GroGPT cos	-0,02	-0,05	-0,04	0,04	-0,05	-0,05
GePpeTto eucl	-0,05	0,09	-0,13	0,01	0,02	0,03
GePpeTto cos	*-0,16	-0,04	*-0,25	-0,04	-0,02	0,00
Minerva eucl	-0,14	0,03	-0,12	-0,09	0,01	-0,14
Minerva cos	0,04	0,05	-0,13	-0,07	0,09	-0,13

Table 9
Spearman correlation between (pseudo)perplexity and human judgment labels, by perturbation type on texts sourced from **Wikipedia**. The asterisk indicates p-value < 0,05.

	(P)PPL ↓		
	no	**swap**	**sub**
mBERT	***-0,21**	***-0,28**	***-0,25**
BERT-ita-xxl	*-0,17	*-0,28	*-0,19
GroGPT	*-0,19	*-0,26	*-0,29
GePpeTto	-0,02	-0,09	-0,12
Minerva	***-0,18**	***-0,32**	***-0,26**

Acknowledgments

This paper is supported by the PRIN 2022 PNRR Project P20227PEPK (EKEEL - Empowering Knowledge Extraction to Empower Learners), funded by the European Union – Next Generation EU, and the LuCET - LingUistic Complexity Evaluation in educaTion - project under the PRIN grant no. 2022KPNY3B funded by the Italian Ministry of University and Research.

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] G. L. Beccaria, Dizionario di linguistica e di filologia, metrica, retorica, Einaudi, 2004.

[3] S. Verberne, L. Boves, N. Oostdijk, P.-A. Coppen, Evaluating discourse-based answer extraction for why-question answering, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 735–736.

[4] B. Elvevåg, P. W. Foltz, D. R. Weinberger, T. E. Goldberg, Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia, Schizophrenia research 93 (2007) 304–316.

[5] D. Iter, J. Yoon, D. Jurafsky, Automatic detection of incoherent speech for diagnosing schizophrenia, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 136–146.

[6] P. Muangkammuen, S. Xu, F. Fukumoto, K. R. Saikaew, J. Li, A neural local coherence analysis model for clarity text scoring, in: Proceedings of the 28th international conference on computational linguistics, 2020, pp. 2138–2143.

[7] S. Gerani, Y. Mehdad, G. Carenini, R. Ng, B. Nejat, Abstractive summarization of product reviews using discourse structure, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1602–1613.

[8] A. Lai, J. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018, pp. 214–223.

[9] F. S. Mim, N. Inoue, P. Reisert, H. Ouchi, K. Inui, Unsupervised learning of discourse-aware text representation for essay scoring, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 378–385.

[10] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, J. Qi, Evaluating document coherence modeling, Transactions of the Association for Computational Linguistics 9 (2021) 621–640.

[11] F. Papa, L. Dini, D. Brunato, F. Dell'Orletta, Unraveling text coherence from the human perspective: a novel dataset for italian, in: Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), 2023.

[12] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4164–4173.

[13] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 997–1006.

[14] L. Pishdad, F. Fancellu, R. Zhang, A. Fazly, How coherent are neural models of coherence?, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6126–6138.

[15] H. C. Moon, M. T. Mohiuddin, S. Joty, C. Xu, A unified neural coherence model, 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, pp. 2262–2272.

[16] M. T. Mohiuddin, P. Jwalapuram, X. Lin, S. Joty, Rethinking coherence modeling: Synthetic vs. downstream tasks, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 3528–3539.

[17] R. Barzilay, M. Lapata, Modeling local coherence: An entity-based approach, Computational Linguistics 34 (2008) 1–34.

[18] P. Laban, L. Dai, L. Bandarkar, M. A. Hearst, Can transformer models measure coherence in text: Re-thinking the shuffle test, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 1058–1064.

[19] D. Iter, K. Guu, L. Lansing, D. Jurafsky, Pretraining with contrastive sentence objectives improves discourse performance of language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4859–4870.

[20] A. Maimon, R. Tsarfaty, A novel computational and modeling foundation for automatic coherence assessment, arXiv preprint arXiv:2310.00598 (2023).

[21] P. Huber, G. Carenini, Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 2376–2394.

[22] S. Duari, V. Bhatnagar, Ffcd: A fast-and-frugal coherence detection method, IEEE Access 10 (2021) 85305–85314.

[23] F. Koto, J. H. Lau, T. Baldwin, Discourse probing of pretrained language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 3849–3864.

[24] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, Discotex at evalita 2023: overview of the assessing discourse coherence in italian texts task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023). CEUR. org, Parma, Italy, 2023.

[25] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy, 2023.

[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 878–891. URL: https://aclanthology.org/2022.acl-long.62. doi:10.18653/v1/2022.acl-long.62.

[27] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 87–94. doi:10.18653/v1/2020.acl-demos.12.

[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[29] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[30] W. de Vries, M. Nissim, As good as new. how to successfully recycle English GPT-2 to make models for other languages, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for

Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 836–846. doi:10.18653/v1/2021.findings-acl.74.

[31] L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, arXiv preprint arXiv:2004.14253 (2020). doi:10.48550/arXiv.2004.14253.

[32] N. Reimers, I. Gurevych, 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, pp. 3982–3992.