Towards an Automatic Evaluation of (In)coherence in Student Essays

Filippo Pellegrino[1,*], Jennifer Carmen Frey[1] and Lorenzo Zanasi[1]

[1]Eurac Research Institute, Viale Druso Drususallee, 1, 39100 Bolzano, Autonome Provinz Bozen - Südtirol

Abstract

Coherence modeling is an important task in natural language processing (NLP) with potential impact on other NLP tasks such as Natural Language Understanding or Automated Essay Scoring. Automatic approaches in coherence modeling aim to distinguish coherent from incoherent (often synthetically created) texts or to identify the correct continuation for a given sample of texts, as demonstrated for Italian in the *DisCoTex* task of EVALITA 2023. While early work on coherence modelling has focused on exploring definitions of the phenomenon, exploring the performance of neural models has dominated the field in recent years. However, coherence modelling can also offer interesting linguistic insights with pedagogical implications. In this article, we target coherence modeling for the Italian language in a strongly domain-specific scenario, i.e. education. We use a corpus of student essays collected to analyse students' text coherence in combination with data perturbation techniques to experiment with the effect of various linguistically informed features of incoherent writing on current coherence modelling strategies used in NLP. Our results show the capabilities of encoder models to capture features of (in)coherence in a domain-specific scenario discerning natural from artificially corrupted texts.

Keywords

Coherence modelling, data perturbation, transformers, education, student essays

1. Introduction

Argumentative essay writing is a fundamental objective in education for both vocational schools and high schools in Italy, as indicated in [1, 2]. It requires students to present arguments supported by personal knowledge or external sources in a coherent and convincing manner. However, writing coherent texts poses both cognitive and linguistic challenges to novice writers and textual competences related to it are frequently claimed to be insufficient, putting pressure on the educational system. Automatically discerning incoherent texts or passages could help teachers to better understand students' problems and give targeted instructions, while students would benefit from more frequent and more timely feedback. However, to date, most NLP research in automatic coherence modelling focused on semantic similarity between two parts of texts using mostly well-formed newspaper or Wikipedia texts, offering little information for educational contexts.

In this study, we explore coherence from an educational perspective, utilizing recent language models and data perturbation techniques to probe their value for linguistically informed and informative automatic coherence evaluation for student essays. While large language models have been used successfully in domain general coherence modelling before, we test their effectiveness for text analysis in this domain-specific scenario, taking into account both surface and non-standard language features. We discuss:

- data perturbation techniques to artificially reproduce real-life scenario incoherence in textual data
- a custom probing task design
- automatic evaluation of coherence using different encoding models

The results of our experiments show the performances of encoder models in recognizing patterns of (in)coherence in a domain-specific educational context such as upper secondary school student essays. The paper is organized as follows: Section 2 provides an overview of previous approaches to coherence modelling and NLP data perturbation with a focus on Italian NLP. Section 3 introduces the data we used for this study, giving information on the research project it originates in as well as on the corpus design and annotation. Section 4 provides a detailed description of our methodology introducing our custom probing tasks (Section 4.1), used Models (Section 4.2.1) and text encoding 4.3 as well as a description of the two analyses performed (Section 4.4 and Section 4.5). Sections 5 and 6 present and discuss our results and Section 7 concludes the article with final considerations.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

filippo.pellegrino.job@gmail.com (F. Pellegrino);
jennifercarmen.frey@eurac.edu (J. C. Frey);
lorenzo.zanasi@eurac.edu (L. Zanasi)

0000-0002-7008-6394 (J. C. Frey); 0000-0002-4439-6567 (L. Zanasi)

2. Related Work

2.1. Coherence modelling

Coherence modeling is an important task in natural language processing (NLP) with potential impact on other NLP tasks such as Natural Language Understanding or automated essay scoring. Early work on coherence modelling focused on the definition of the phenomenon [3, 4, 5, 6, 7] and provides valuable frameworks such as Centering Theory [8, 9] and Entity-Grid approach [10]. Following the great development of neural network systems in recent years, many works such as [11, 12, 13, 14] explored coherence modelling implementing further and more sophisticated solutions for the English language. Recently, the Italian NLP community has approached the topic from an engineering point of view, using Italian pre-trained neural models to distinguish coherent from (mainly synthetically constructed) non-coherent texts [15, 16, 17, 18]. Some efforts were also made for multilingual scenarios [19] demonstrating the encoding capabilities of multilingual models for coherence features.

2.2. Data perturbation

In data perturbation, dataset entries are corrupted with specific computational operations to simulate noise condition and test the model performance on real world conditions [20]. Many studies on data perturbation and data augmentation in NLP focus on model agnostic methods [20, 21, 22, 23] using random deletion, random swap, synonym replacement, random insertion and punctuation insertion techniques for text classification with limited amount of data. More sophisticated and task-oriented data augmentation approaches are proposed for sentiment analysis [24], hate speech classification [25], hypernymy detection [26] and domain specific classification [27].

3. Data

The data used in this study originates from a research project, conducted in South Tyrol between 2020 and 2024. The project named ITACA: Coerenza nell'ITAliano Accademico [28] had the aim to study textual competences of students in their first language Italian with particular focus on aspects of text coherence. Within the project various outcomes have been produced: a corpus of Italian student essays collected in Italian South Tyrolean upper secondary schools, a validated rating scale to evaluate coherence in student essays, and coherence ratings for texts in the corpus from three independent raters using the previously developed rating scale. The products are described in the following section.

3.1. ITACA Corpus

The ITACA corpus[1] is an annotated learner corpus created within the project ITACA: Coerenza nell'ITAliano Accademico [28]. It consists of a total of 636 argumentative essays from Italian L1 upper secondary school students from the autonomous province of Bolzano/Bozen[2] during the school year 2021/2022. The texts were collected by asking 12th grade students to type an argumentative essay following precise indications of writing time, text length and topic. The full assignment can be consulted in the Appendix B. While the assignment asked for a minimun text length of 600 words, the average number of tokens in the essay is with 668, just slightly above the minimum length requirement.
The totality of the 636 collected texts constitutes 382,964 tokens. All data were collected digitally and anonymously and underwent subsequent control and cleaning procedures, partly manually, to ensure their integrity and to guarantee the anonymity of the participants. Essays were collected, by asking students to type their essays into an input field in an online form, additional metadata was collected by a subsequent online questionnaire asking for basic socio-demographic information, students' language background, and reading and writing habits. The whole corpus was automatically tokenized, lemmatized and annotated for part-of-speech and syntactic dependencies with the support of project collaborators from Fondazione Bruno Kessler, who also supported the project in the setup of an interface for manual annotation based on Inception[29].

A manual annotation of a subset of 388 texts was performed by two trained annotators and offers detailed descriptions of the text's structure, with a focus on the use of various linguistic features (such as punctuation, connectives, agreements, anaphora, contradictions) that enhance or limit the text's cohesion and coherence.
The manual annotation of the corpus was guided by the three sections elaborated in [30] and contained annotations for traits of incoherence referring to

1. segmentation (e.g. splice comma, added comma, not-signed parenthetical clause)
2. logic-argumentative plan (e.g. issues in the use of connectives, contradictions)
3. thematic-referential plan (e.g. critical agreement, critical anaphora, not-expanded comment)

The corpus is accessible through an ANNIS search interface [3]and can be downloaded in various formats from the Eurac Research Clarin Center (ERCC) under the CLARIN ACADEMIC END-USER LICENCE ACA-BY-NC-NORED

[1] https://www.porta.eurac.edu/lci/itaca/
[2] texts are collected in Bolzano, Bressanone, Merano and Brunico
[3] https://commul.eurac.edu/annis/itaca

1.0 licence [4]. Downloads and further documentation can also be accessed via Eurac Research's PORTA platform[5].

3.2. Manual coherence ratings

Each single essay was additionally manually evaluated in a double-blind manner by a panel of six experts who applied a specially created, rating scale, which was subsequently validated to assess textual coherence. The items were rated on a Likert scale from one to ten and referred to three dimensions of coherence (structure, comprehensibility, segmentation). The average structure score μ is attested at 4.55 with standard deviation σ = 5. For comprehensibility, μ = 6.29 and σ = 1.65, while for segmentation μ = 5.99 and σ = 1.79.

4. Methodology

In this study, we focus on NLP data perturbation [20, 21] and custom probing tasks [31] to evaluate the ability of Italian BERT models of discerning features of coherence given different pre-training conditions and fine tuning. In our analysis, we aim to evaluate automatic coherence modelling techniques, applying them to student essays with varying degrees of well-formedness and coherence. We conducted a number of experiments probing whether state-of-the-art coherence modelling techniques based on BERT encodings would be able to distinguish between original, i.e. allegedly coherent texts and those containing features of incoherence identified for student writing before. In our case study, we use data perturbation techniques to reproduce specific students' errors observed during the textual analysis of the ITACA project [28] (see Section 3), in order to apply text modification in a fully controlled fashion. We used representations obtained from BERT [32] models to demonstrate the ability of automatic systems to encode patterns of (in)coherence in a specialized scenario such as Italian student essays and evaluate their potential for educational purposes.

4.1. Custom Probing Tasks

Using data perturbation techniques, we aim to reproduce both general-purpose coherence modelling perturbation strategies and modifications inspired by some of the most salient features of textual (in)coherence observed in the annotation process for the ITACA project. These include incoherent order of arguments and sentences, incorrect use of connectives, overuse of polyfunctional connectives, unresolved co-reference, the use of splice comma and an overuse of paratactical constructions. Assuming that students would not produce the these features throughout the whole essay, but only struggle occasionally (e.g. not all connectives are semantically incorrect), we reduced the perturbation ratio to 50% in Pronoun Perturbation, Splice Comma Perturbation and Parataxis Perturbation in order to create realistic conditions and increase the difficulty of the single tasks. Although data perturbation can also operate on the character level, we opted for token- and sentence-level approaches maintaining parameters in a controlled setting.

We implemented the following custom probing tasks:

Sentence Order Perturbation [SHUFF]:
As in other synthetic datasets for coherence modelling [15] this data perturbation technique is to randomly shuffle sentences within the texts.

Connective Perturbation [LICO]:
In order to imitate texts in which the logical connection between phrases is erroneous, we randomly substituted connectives used in the text exploiting both manual and automatic processing with Stanza[6]; To identify the connectives to substitute, we referred to a string matching of all connectives listed in the Lexicon of Italian Connectives (LICO) [33].

Polyfunctional Connective Perturbation [POLYFUNCT]:
Based on the ITACA corpus annotation scheme, we implement a probing task, imitating young writers tendency to use simple polifunctional connectives instead of highly semantically loaded ones. For this, we substitute all connectives in the text by the polyfunctional connective "e".

Pronoun Perturbation [PRON]:
For a very simplistic approximation of corrupted anaphoric references, we identified pronouns with Stanza and replaced them randomly by other pronouns isoleted from the corpus. To ensure a minimum of correct pronouns, only 50% of the pronouns in the text were corrupted.

Splice Comma Perturbation [SPLICE]:
A splice comma is the use of a comma to join two independent sentences. The comma can substitute a dot, a colon, or semicolon [34, 35, 36, 37]. In our case, long pause markers such as periods, colons, or semicolons were substituted with a comma. We apply the perturbation to just 50% of the conjunctions in the text to partially keep punctuation unaltered.

[4] http://hdl.handle.net/20.500.12124/76
[5] https://www.porta.eurac.edu/itaca
[6] https://stanfordnlp.github.io/stanza/

Perturbation	Example Sentence
None	Stamattina io sono andato al mercato. Ho comprato delle mele e delle arance. Poi sono tornato a casa e ho preparato una torta.
Sentence Order Perturbation	Poi sono tornato a casa e ho preparato una torta. Stamattina io sono andato al mercato. Ho comprato delle mele e delle arance.
LICO Connective Perturbation	Stamattina io sono andato al mercato. Ho comprato delle mele e delle arance. Poi sono tornato a casa invece di ho preparato una torta.
Polyfunctional Connective Perturbation	Stamattina io sono andato al mercato. Ho comprato delle mele e delle arance. e sono tornato a casa e ho preparato una torta.
Pronoun Perturbation	Stamattina noi sono andato al mercato. Ho comprato delle mele e delle arance. Poi sono tornato a casa e ho preparato una torta.
Splice Comma Perturbation	Stamattina io sono andato al mercato, Ho comprato delle mele e delle arance, Poi sono tornato a casa e ho preparato una torta.
Parataxis Perturbation	Stamattina io sono andato al mercato. Ho comprato delle mele, delle arance. Poi sono tornato a casa. ho preparato una torta.

Table 1
Example Sentences under Text Perturbations. The example corresponds to the English "This morning I went to the market. I bought some apples and oranges. Then I went back home and baked a cake"

Parataxis Perturbation [PARATAX]:
Coordinating conjunctions extracted with Stanza are substituted with punctuation taken from a list to create paratactic sentences. We apply the perturbation to just 50% of the conjunctions in the text to keep some conjunctions untouched.

Text perturbation examples can be consulted in Table 1

4.2. Models

4.2.1. Pre-trained Models

For our experiments, we test three different BERT-based models to obtain vector representations for our probing tasks.

1. BERT-ita base [38]: trained with Italian data from the OPUS corpora collection[7] and Wikipedia[8].The final training corpus has a size of 13GB and 2,050,057,573 tokens.
2. GilBERTo[9]: RoBERTa based model [39]. The model is trained with the subword masking technique for 100k steps managing 71GB of Italian text with 11,250,012,896 words [40]. The team took up a vocabulary of 32k BPE subwords, generated using SentencePiece tokenizer [41].

4.2.2. BERT-ita Fine-tuning

Inspired by the works of [42] and [43], the BERT-ita model was fine-tuned using a dataset of high school essays typologically similar to our dataset, thankfully provided for this purpose by the Fondazione Bruno Kessler (FBK). The number of essays employed for the fine-tuning corresponds to 2096 dataset entries with a mean text length of 705 tokens. Fine-tuning our BERT model allowed us to provide further contextual and text essay style information to the pre-trained model, increasing the model's ability in domain-specific text representation. The provided hyperparameter configuration for training is: truncation = max length, padding = max length, batch size = 16, learning rate = 5e-5 and epochs = 2. The model is trained on both *Masked Language Modeling* and *Next Sentence Prediction* tasks [32]. Taking into account the limited amount of data and the relatively quick training time, we use the L4 GPU available in Google Colab[10] (pro version).

4.3. Text Encoding

We retrieved vector representations and performed a binary text classification experiment for each perturbation technique[11]. The model is fed with batch size = 1 with all the texts contained in the set. To overcome the length input limit of 512 tokens imposed by BERT models and process the entire text in a row with no loss of contextual information, we split the text into two segments when reached the max input lenght. Furthermore, we adopted a mean-pooling strategy by calculating the mean between the last hidden state of each contextualized token embedding in the batch across the input sequence length. The final text representation is the mean of all segment embeddings in the batch.

[7] https://opus.nlpl.eu/
[8] https://it.wikipedia.org/wiki/Pagina_principale
[9] https://github.com/idb-ita/GilBERTo?tab=readme-ov-file
[10] https://colab.research.google.com/
[11] The code for this part of the project was written with the help of the AI tool Chat GPT.

4.4. Model Performance Analysis

We first perform a model performance analysis, comparing the model performance in classification for each of the custom probing tasks with each of the three models. Classification is performed with a Random Forest classifier [44], defining each experiment as a binary classification between the original and perturbated texts. The classes were balanced across the entire dataset. To optimize the amount of available data for training and testing, we use 10-fold cross-validation for evaluation. We compare model performance against a majority class baseline (0.5 for balanced binary classification) and against each other using f1 scores.

4.5. Error Analysis

In a subsequent analysis, we compare the model predictions of our best-performing model with the human coherence ratings provided for the corpus. In order to obtain a single coherence score for each essay, the scores were averaged over the different annotators and the three components (structure, comprehensibility and segmentation; see Section 3). We perform an error analysis by comparing the predictions for unmodified texts with the highest and lowest coherence scores using a random forest classifier trained with the model that achieved the best results in the model comparison. Assuming that all tasks have the same weight, we select the best performing model according to the average f1 score achieved in the model performance analysis (see Section 4.4). The train set for this evaluation corresponds to 90% of the data, while the test set represents the 5% of essays with the highest (μ = 8.28, σ = 0.36) and the 5% with the lowest coherence scores (μ = 2.63, σ = 0.51). Finally, we interpret the results, manually investigating texts that were misclassified as modified texts from both tails of the test set.

5. Results

The classification experiments show the ability of the BERT models to encode the features of (in)coherence represented by the perturbation techniques introduced in Section 4.1. The following sections illustrate our findings for the BERT model comparison and the error analysis conducted on a selected subset of non-modified texts.

5.1. Models Comparison Analysis

F1 scores for most models were very similar with just small differences between the three models. In average, GilBERTo was found to be the best performing model for most tasks, probably due to its higher amount of training data and its lighter model architecture. However, we do not expect these differences to be significant. Except for the improvement in the shuffling task after fine-tuning, the ITACA-bert model remains comparable to its base version, probably due to the scarcity of domain-specific training data. Results showed that models achieved better performance on semantic tasks such as polyfunctional conjunction perturbation or pronoun perturbation while struggling with syntactic probing tasks such as shuffling and splice comma perturbation. For the shuffling task, a considerable improvement can be observed after fine-tuning (+0.12% from F1 = 0.38 to F1 = 0.50). However, neither of the shuffling models performs better than a random baseline, while the splice comma experiment models performed slightly better, with the BERT-ita and Gilberto models marginally beating the baseline of 0.5. A graphical comparison between model performances can be seen in Figure 1.

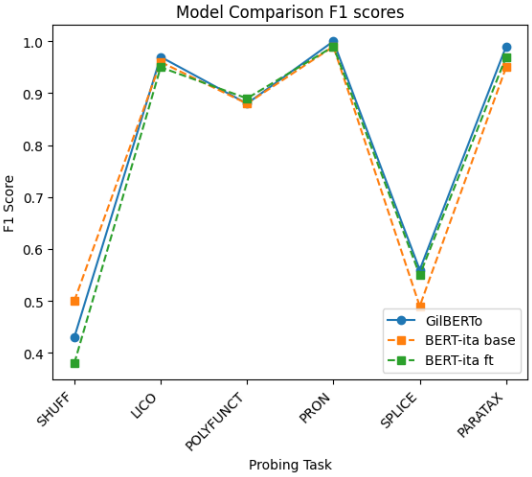

Figure 1: Model performances comparison on single probing tasks

A detailed overview of the classification results for single tasks and models can be found in the Appendix A. The tables provide measures of the f1 score for each experiment and model.

5.2. Error analysis on evaluation set

To better observe the encoding and classification performance of BERT, we decide to isolate the texts with the highest and the lowest coherence scores according to the average coherence scores as specified in 4.5. The resulting test set corresponds roughly to the 10% of the total number of texts in the corpus. Our expectation is that texts with lower coherence scores have a higher chance to be misclassified as modified texts, while texts with higher coherence scores should not lead the classifiers to identify traits of incoherence as specified in the cus-

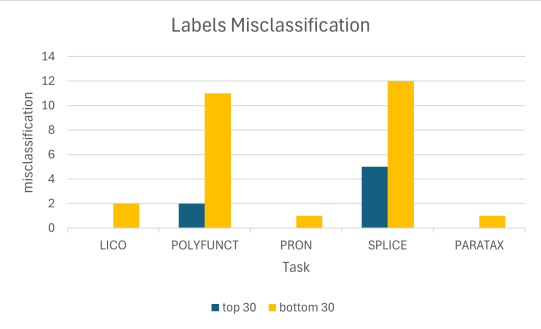

Figure 2: Classification results on evaluation set. The figure shows the amount of misclassified labels for the essays that lie in the highest and lowest tail of the score ranking ITACA dataset.

tom probing tasks. We perform all analysis using the GilBERTo model for text encoding, as it was revealed to be the best performing model when averaging f1 scores on all tasks of the model performance analysis (see Section 4.4). However, we exclude the shuffling task as model performance was below the baseline and therefore too low for interpretation. Thus, we train a random forest classifier with the 90% of the train set, for all custom probing tasks described in Section 4.1.

Our results show that the distribution of misclassified labels is generally skewed toward texts with lower coherence scores, but misclassifications for texts with higher coherence scores were also found. While the splice comma and polyfunctional conjunction (see Figure 2) probing tasks showed clearly more misclassifications on the lower tail of the dataset, also well-rated texts were occasionally misclassified as perturbed texts. On the contrary, the small number of misclassifications on the parataxis and pronoun perturbation probing tasks might suggest that the operationalizations taken in this work are too simplistic to be representative of students' mistakes in the texts and, therefore, not able to pick up on traits of incoherence present in the students' essays. The results of the experiment can be consulted in Appendix A.

6. Discussion

Although data perturbation cannot fully reproduce the variability of real-word students' mistakes, our results give precious insights about the ability of BERT encoders to capture degrees of coherence on both syntactic and semantic level. Of course, the efficiency of the data perturbation might be influenced by several factors, such as the fact that the original texts used for our experiments already naturally contain errors of the same or other types. However, we argue that this is the case for any type of data set of unknown quality that is subject to automatic coherence evaluation. Thus, before the evaluation, texts have not been subjected to any review and, excluding other external factors, they reproduce real-world writing conditions. The results of language encoding and classification depend on the difficulty of the perturbation task and on the original training of the BERT model. However, despite the fact that the BERT-ita base and GilBERTo exploit different training strategies, no drastic performance fluctuations have been observed on our selected language tasks. Even though the effects of fine-tuning with domain-specific data is limited to the amount of affordable data, the effect can already be observed by looking at the increment on the shuffling task performance.

The classification of the evaluation set highlighted the potential of data perturbation techniques for the encoding of (in)coherence features. Previous approaches to coherence modelling implemented solutions inspired by theoretical intuitions. In our case, we decided to start from natural textual errors and check the ability of the model in capturing the same features presented in the text. For a more transparent interpretation of results and explanation of individual classification it would be of interest to check how attention maps change according to the tuning of the model [45].

7. Conclusion

In this paper, we presented an evaluation of coherence modelling techniques for detecting incoherence in student essays based on surface-level features of incoherence. We used the ITACA corpus of Italian upper secondary school essays to perform a number of classification techniques using data perturbation and BERT-based text encoding methods. After a preliminary comparison between pre-trained and fine-tuned models we adopted the best performing one according to our results. The results of the chosen tasks are influenced by the implementation of the perturbation technique, the encoding ability of the model, and the amount and the quality of the data the model is pre-trained on. The best performances are bounded to the model pre-trained with the highest amount of data (GilBERTo). We based our evaluation on simple f1 measures considering this sufficiently indicative of the encoding ability of the model applied to each specific probing task.

Since we mainly tested custom perturbation techniques and the encoding abilities of BERT models, future research directions might involve data perturbation techniques enhancement, XAI techiques for model behaviour analysis [46, 45] and the exploitation of state-of-the-art generative one shot and few-shot models in a highly domain-specific scenario such as school essays writing.

Acknowledgments

We thank Fondazione Bruno Kessler Trento for their support on the ITACA corpus and for allowing us to use their student essay dataset for fine-tuning.

References

[1] d. e. d. R. Ministero dell'Istruzione, Indicazioni nazionali per i licei, Ministero dell'Istruzione, dell'Università e della Ricerca, Roma, Italia, 2010.

[2] d. e. d. R. Ministero dell'Istruzione, Istituti tecnici: linee guida per il passaggio al nuovo ordinamento, Ministero dell'Istruzione, dell'Università e della Ricerca, Roma, Italia, 2010.

[3] T. A. Van Dijk, Context and cognition: Knowledge frames and speech act comprehension, Journal of pragmatics 1 (1977) 211–231.

[4] T. Reinhart, Conditions for text coherence, Poetics today 1 (1980) 161–180.

[5] F. Danes, Functional sentence perspective and the organization of the text, Papers on functional sentence perspective 23 (1974) 106–128.

[6] P. H. Fries, On the status of theme in english: Arguments from discourse, Micro and macro connexity of texts 45 (1983).

[7] J. R. Hobbs, Coherence and coreference, Cognitive science 3 (1979) 67–90.

[8] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: a framework for modelling the coherence of discourse (1994).

[9] B. Di Eugenio, Centering in italian, arXiv preprint cmp-lg/9608007 (1996).

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

[11] Y. Farag, H. Yannakoudakis, T. Briscoe, Neural automated essay scoring and coherence modeling for adversarially crafted input, arXiv preprint arXiv:1804.06898 (2018).

[12] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 4328–4339.

[13] J. Li, E. Hovy, A model of coherence based on distributed sentence representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 2039–2048.

[14] D. T. Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1320–1330.

[15] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, et al., Discotex at evalita 2023: overview of the assessing discourse coherence in italian texts task, in: CEUR WORKSHOP PROCEEDINGS, volume 3473, CEUR, 2023, pp. 1–8.

[16] M. Galletti, P. Gravino, G. Prevedello, Mpg at discotex: Predicting text coherence by treebased modelling of linguistic features, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, 2023.

[17] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme (2022).

[18] E. Zanoli, M. Barbini, C. Chesi, et al., Iussnets at disco-tex: A fine-tuned approach to coherence, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, 2023.

[19] D. Brunato, F. Dell'Orletta, I. Dini, A. A. Ravelli, Coherent or not? stressing a neural language model for discourse coherence in multiple languages, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 10690–10700.

[20] M. Moradi, M. Samwald, Evaluating the robustness of neural language models to input perturbations, arXiv preprint arXiv:2108.12237 (2021).

[21] Y. Zhang, L. Pan, S. Tan, M.-Y. Kan, Interpreting the robustness of neural nlp models to textual perturbations, arXiv preprint arXiv:2110.07159 (2021).

[22] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, arXiv preprint arXiv:1901.11196 (2019).

[23] A. Karimi, L. Rossi, A. Prati, Aeda: an easier data augmentation technique for text classification, arXiv preprint arXiv:2108.13230 (2021).

[24] H. Q. Abonizio, E. C. Paraiso, S. Barbon, Toward text data augmentation for sentiment analysis, IEEE Transactions on Artificial Intelligence 3 (2021) 657–668.

[25] G. Rizos, K. Hemker, B. Schuller, Augment to prevent: short-text data augmentation in deep learning for hate-speech classification, in: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 991–1000.

[26] T. Kober, J. Weeds, L. Bertolini, D. Weir, Data augmentation for hypernymy detection, arXiv preprint arXiv:2005.01854 (2020).

[27] T. Nugent, N. Stelea, J. L. Leidner, Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation, in: Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14,

Springer, 2021, pp. 157–169.
[28] A. Bienati, C. Vettori, L. Zanasi, In viaggio verso itaca: la coerenza testuale come meta della scrittura scolastica. proposta di una griglia di valutazione, Italiano a scuola 4 (2022) 55–70.
[29] J.-C. Klie, M. Bugert, B. Boullosa, R. E. De Castilho, I. Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th international conference on computational linguistics: System demonstrations, 2018, pp. 5–9.
[30] A. Ferrari, Linguistica del testo. Principi, fenomeni, strutture, volume 151, Carocci, 2014.
[31] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni, What you can cram into a single vector: Probing sentence embeddings for linguistic properties, arXiv preprint arXiv:1805.01070 (2018).
[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[33] A. Feltracco, E. Jezek, B. Magnini, M. Stede, Lico: A lexicon of italian connectives, CLiC it (2016) 141.
[34] C. E. Roggia, Una varietà dell'italiano tra scritto e parlato: la scrittura degli apprendenti, Ferrari A., De Cesare AM (2010) (2010) 197–224.
[35] L. Cignetti, Didattica della scrittura e linguistica del testo: tre priorità di intervento, Ostinelli M.(a cura di), La didattica dell'italiano. Problemi e prospettive, DFA SUPSI, Locarno (2015) 14–24.
[36] A. Colombo, A me mi. Dubbi, errori, correzioni nell'italiano scritto: Dubbi, errori, correzioni nell'italiano scritto, FrancoAngeli, 2010.
[37] M. Prada, Scritto e parlato, il parlato nello scritto. per una didattica della consapevolezza diamesica, Italiano LinguaDue 8 (2016) 232–260.
[38] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.
[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[40] J. Abadji, P. O. Suarez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, arXiv preprint arXiv:2201.06642 (2022).
[41] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv preprint arXiv:1808.06226 (2018).
[42] D. Licari, G. Comandè, Italian-legal-bert: A pretrained transformer language model for italian law., EKAW (Companion) 3256 (2022).
[43] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[44] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[45] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does bert look at? an analysis of bert's attention, arXiv preprint arXiv:1906.04341 (2019).
[46] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, arXiv preprint arXiv:2010.00711 (2020).

A. Appendix A

Aug Techniques	GilBERTo F1 Score	ITACA-bert F1 Score	BERT-base-italian F1 Score
SHUFF	0.43	0.5	0.38
LICO	0.97	0.96	0.95
POLYFUNCT	0.88	0.88	0.89
PRON	1.0	0.99	0.99
SPLICE	0.56	0.49	0.55
PARATAX	0.99	0.95	0.97

Table 2
Model comparison on f1 score for each task. Each probe is run as a binary classification task on 636 dataset entries. The baseline is set on 0.5

Aug Techniques	Train Dataset Len	Num Labels	Baseline	Accuracy
LICO	575	2	0.5	0.96
POLYFUNCT	575	2	0.5	0.78
PRON	575	2	0.5	0.98
SPLICE	575	2	0.5	0.7
PARATAX	575	2	0.5	0.98

Table 3
Error analysis

B. Appendix B

"In base all'esperienza maturata durante la pandemia di Covid-19, il Ministro dell'Istruzione ha proposto di estendere permanentemente, a partire dal prossimo anno scolastico, la Didattica Digitale Integrata (DDI, modalità didattica che combina momenti di insegnamento a distanza e attività svolte in classe) al triennio delle scuole superiori [...]. Immagina di dover scrivere una lettera al Ministro in cui esponi le tue ragioni a favore o contro questa possibilità, argomentandole in modo da convincerlo della bontà delle tue idee [...]. Durante lo svolgimento del testo ricordati di: 1. Chiarire la tesi che intendi difendere. 2. Spiegare le motivazioni a sostegno della tesi. 3. Prendere in considerazione il punto di vista alternativo e illustrare le ragioni per cui non sei d'accordo. 4. Arrivare a una conclusione. 5. Prima di consegnare, ricordati di rileggere con cura il testo che hai scritto. Il tuo obiettivo è convincere il Ministro della bontà della tesi che sostieni. Hai 100 minuti di tempo per scrivere un testo di almeno 600 parole."