DisCoTex at EVALITA 2023: Overview of the Assessing DIScourse COherence in Italian TEXts task

Dominique Brunato[1], Davide Colla[2], Felice Dell'Orletta[1], Irene Dini[1,3], Daniele Paolo Radicioni[4] and Andrea Amelio Ravelli[5]

[1]*Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa - ItaliaNLP Lab*
[2]*Dipartimento di Studi Storici - Università degli Studi di Torino*
[3]*University of Pisa*
[4]*Dipartimento di Informatica - Università degli Studi di Torino*
[5]*Dipartimento di Lingue, Letterature e Culture Moderne - Università di Bologna*

Abstract

The Assessing DIScourse COherence in Italian TEXts (DisCoTeX) task is the first shared task focused on modelling discourse coherence for Italian real-word texts, which has been proposed for the first time at EVALITA 2023. Providing two different datasets from different textual genres, we arranged the task into two independent tasks: a more traditional one, aimed at evaluating whether models are able to distinguish well-organized documents from corrupted ones and a less explored one, which assesses the models' performance on texts evaluated for coherence by human raters. In this paper, we describe the datasets released, we discuss the different approaches tackled by the participating systems and provide a first analysis of the obtained results.

Keywords

text coherence, Italian language, computational modeling, evaluation campaign, dataset

1. Introduction and Motivation

Coherence is a key property of any well-organized text and it plays a crucial role in human discourse processing. Indeed, as individuals process unfolding text, they are required to assemble information from single sentences and to draw inferences between and among them in order to create a meaningful mental representation of the whole text. According to the tripartite model developed by [1], this is the outcome of a three-step process in which readers construct multileveled memory representations of a text, encoding different, and progressively more abstract, information at each level. From this perspective, coherence is an inherently psychological construct, thus very hard to be modelled; however, it also has a counterpart at the level of linguistic content and structure, often referred to as "cohesion", a property of a text that is conveyed by signalling linguistic devices such as reference, ellipsis, discourse connectives, argument overlap, which help readers make explicit the logical links between different units in texts.

As regards the computational modelling, coherence has been widely investigated in the Natural Language Processing (NLP) community, particularly in the "pre-deep-learning" era, with much research drawing inspiration from frameworks like the Centering Theory [2]. One popular approach in this context is the entity-grid approach, which focuses on assessing *local* coherence, specifically the transitions between adjacent sentences (see, among others, [3, 4]). More recently, also neural models have been applied to deal with both structured representations of text and unstructured text by taking advantage of neural models' ability to learn useful representations for the task, e.g. [5, 6]. Modelling coherence in natural language is of pivotal importance in a variety of downstream applications, from automatic essay scoring in language learning scenarios [7, 8], to language assessment in clinical settings [9, 10] Additionally, from the Natural Language Generation point of view, coherence is an intrinsic evaluation metric to assess the quality of generated texts. An emerging area of interest pertains to the interpretability of modern deep neural networks. In this respect, while existing work on probing pre-trained language models has largely focused on sentence-level properties, the ability of these models to encode discourse and pragmatic phenomena is still unclear [11, 12, 13].

Recognizing the fundamental role that coherence plays across a variety of scenarios and the challenges in developing a unified metric to quantify this concept, DisCoTeX, organized in the context of the 8th evaluation campaign of NLP and speech tools for the Italian language (EVALITA 2023) [14] intends to encourage research on automatic discourse coherence modeling with empha-

EVALITA 2023: 8th Evaluation Campain of NLP and Speech Tools for Italian, Semptember 07–08, 2023, Parma, Italy
dominique.brunato@ilc.cnr.it (D. Brunato); davide.colla@unito.it (D. Colla); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); irene.dini@ilc.cnr.it (I. Dini); daniele.radicioni@unito.it (D. P. Radicioni); andreaamelio.ravelli@unibo.it (A. A. Ravelli)

sis on the Italian language.

2. Definition of the Task

Drawing inspiration from existing coherence modeling literature, the DisCoTeX task was designed with the intention of addressing two distinct scenarios. The first scenario involves the evaluation of models' ability to differentiate well-structured documents from corrupted ones. The corrupted documents are typically created by either shuffling the sentence order of the original document or replacing specific linguistic elements that contribute to coherence within and across sentences, such as personal pronouns or discourse connectives. The second scenario, which has been less explored, focuses on assessing the models' performance in coherence evaluation by comparing their predictions to human raters' evaluations.

To capture these distinct scenarios, we proposed two independent sub-tasks:

- Sub-task 1 - **Last sentence classification**: This sub-task was casted as a binary classification task. Specifically, participants are presented with a prompt, which is a short paragraph consisting of approximately three consecutive sentences, and an individual sentence referred to as the target. The objective is to classify whether the target sentence, when combined with the prompt, forms a coherent or incoherent text. The negative target can either be a sentence randomly selected from a different document or a sentence extracted from the same document as the prompt, in order to introduce incremental degrees of complexity on the resolution of the task;
- Sub-task 2 - **Human score prediction**: This subtask was framed as a regression task where participants were asked to predict the average coherence score assigned by human raters to short paragraphs. These paragraphs were evaluated in their original or artificially modified version. As shown in previous tasks on the automatic assessment of subjective phenomena [15, 16], this scenario is expected to be more challenging, as it requires modeling the human perception of coherence, which can be influenced by both linguistic and non-linguistic factors, as highlighted in previous studies [7].

For both sub-tasks, dataset were extracted from two corpora representative of two distinct domains, as described in the following section.

3. Datasets

The dataset[1] utilized for the DisCoTeX task encompasses texts sourced from two distinct origins: the Italian Wikipedia and the Italian speech transcripts section of the Multilingual TEDx corpus (mTEDx). These sources represent two different language varieties: the former is a 'standard' written variety, and the latter a 'hybrid' variety combining diverse genres (e.g., university lectures, newspaper articles, conference presentations, and TV science programs) as well as different semiotic modes, such as written, spoken, audio, and video [17]. Extensive research on genre and register variation acknowledges that written and spoken language employ distinct strategies to establish coherence within a text [18]. Therefore, we decided to evaluate systems on both these types of data.

For sub-task 1, each data sample consists of a prompt, which is a paragraph comprising three sentences, followed by a target sentence. To create the written dataset, we leveraged the existing paragraph segmentation in Wikipedia to select four-sentence paragraphs. For the spoken dataset, as mTEDx speeches lacked such internal structure, we divided all the transcripts into passages of four sentences. The target sentence is determined as the immediate continuation of the prompt, forming a coherent sample. In the case of a non-coherent passage, as previously anticipated, we selected either a sentence randomly taken from a different document or the sentence that appears ten sentences after the prompt in the same document. Each final dataset consists of 8,000 training samples and 800 test samples. Examples can be found in Table 1.

Regarding sub-task 2, the dataset construction differs slightly. In this case, for each source we extracted samples consisting solely of four-sentence paragraphs (we keep the term 'prompts' to refer to them), with half of them deliberately made incoherent through sentence perturbations. The possible perturbations, chosen with equal probability, include:

- Flip of two random sentences: each sentence of the prompt has the same probability of being flipped.
- Swap of a sentence with the next 10th of the same document from which the prompt was extracted. The first and the last sentence have double the swap probability compared to the middle two sentences to make the swapping of the first/last or a middle one equiprobable.

For the purposes of the DisCoTeX task, we selected 1,064 prompts equally balanced between the two domains.

[1]The DisCoTeX dataset is available at the following link: https://github.com/davidecolla/DisCoTex/

Prompt	Target	Class
Il regolamento del carcere era durissimo e le condizioni igieniche drammatiche. Agli ebrei erano negati i pochi diritti concessi agli altri prigionieri politici e comuni, ovvero l'ora d'aria in cortile, l'assistenza sanitaria, la possibilità di ricevere lettere e pacchi e di acquistare generi alimentari allo spaccio del carcere. Gli interrogatori degli arrestati erano condotti in uno stanzone a pian terreno, detto il "refettorio".	Qui le sevizie di ogni genere venivano inflitte soprattutto sugli ebrei che non rivelavano i recapiti o i nascondigli dei loro parenti della cui presenza a Milano o nei dintorni le SS erano venute a conoscenza tramite loro spie.	1
Ci siamo trovati a Brasilia, la capitale del Brasile; e c'erano città di tutto il mondo, dall'Australia al Giappone, all'Asia, all'Africa, agli Stati Uniti. E lì abbiamo avuto la consapevolezza che siamo un movimento che sta crescendo nel mondo e che sempre più costruisce risultati e vantaggi. Una delle più grandi città del mondo che ha fatto questa scelta è San Francisco.	Vedete in questo semplicissimo grafico, il rosso è tutto quello che prima,una decina di anni fa, andava a smaltimento.	0

Table 1
Examples extracted from the dataset of sub-task 1. The first one belongs to the Wikipedia dataset and the second one to mTEDx.

Of these, 33% (i.e. 360 prompts) are extracted from the subset of authentic prompts and 66% (i.e. 704) from perturbed ones. Examples can be found in Table 2.

As anticipated, coherence was assessed through manual annotation. Specifically, to gather human ratings of coherence, we conducted a crowdsourcing task on the Prolific[2] platform, involving Italian native speakers. Recognizing that coherence is a subjective concept influenced by the reader or listener's interpretation, we employed a gradual judgment approach, and asked them to evaluate their perception of on a Likert scale ranging from 1 to 5. The number of annotations per prompts ranged from 9 to 12, with an average of 11.75.

The resulting dataset was split into training and test samples with a proportion of 80% to 20%, respectively.

In Figure 1 we show some general statistics about collected judgments considering both the whole dataset of prompts and prompts grouped into specific sections, according to genre and perturbation. As it can be seen, prompts derived from Wikipedia texts are generally rated as more coherent by humans compared to TEDx prompts. This observation confirms previous findings with regard to the influence of genre on the perception of coherence [7]. What is particularly interesting is that this disparity is evident not only in the original form of the prompts but also in the perturbed versions. This seems to suggest that Wikipedia documents tend to exhibit a more standardized structure, including internal coherence, which remains relatively stable even with minor alterations that affect sentence order or the insertion of an intruder sentence from the same document. We plan to conduct a more in-depth analysis by examining each perturbation strategy independently to gain a deeper understanding of their individual impact on coherence.

[2]https://www.prolific.co/

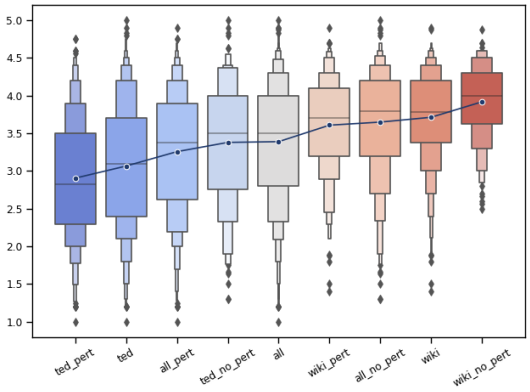

Figure 1: Overview of human judgments collected for the dataset used in sub-task 2. The plot shows the overall mean of human judgments for the whole dataset (*all*) and for respective subsets, including both coherent (**_no_pert*) and perturbed prompts (**_pert*).

3.1. Format

The DisCoTEX dataset was released as tab-separated text files. Specifically, for sub-task 1, the two data sources (i.e. Wikipedia and TED) were kept separated and, for each source, participants were provided with a file with a following structure:

- ID: a numerical identifier for the entry;
- PROMPT: textual passage made by three consecutive sentences;
- TARGET: the sentence which participants are asked to assess if it is coherent with the prompt (i.e. it is the next sentence after the prompt);

Text	Mean	Stdev
Le nuove idee sono una sfida che accende nel nostro cervello la stessa area che elabora le minacce fisiche. Ecco perché tendiamo a reagire con forza, a volte con aggressività, alle nuove idee. Davanti a informazioni che mettano in discussione le nostre convinzioni noi tendiamo paradossalmente a reagire rafforzandole ancora di più. Si chiamano bias cognitivi, sono molto forti e ci caschiamo tutti.	4.83	0.58
I Romani furono scommettitori appassionati, specialmente ai tempi dell'Impero Romano, e il gioco dei dadi era popolare, seppur proibito da una "Lex alearia" del 204 a.C. circa, eccetto che durante i Saturnali. Orazio derise la gioventù dell'epoca che sprecava tempo tra i pericoli del gioco invece di domare il suo cavallo e darsi alle durezze dell'inseguimento. Una di queste diceva che nessuna causa poteva essere intentata da una persona che permetteva il gioco d'azzardo nella sua casa anche se era stata imbrogliata o assalita. Le scommesse sui dadi per denaro fu l'oggetto di molte leggi Romane.	3.3	0.95

Table 2
Examples extracted from the dataset of sub-task 2. The first one is an original prompt taken from the mTEDx corpus. The second ones is a perturbed prompt from the Wikipedia corpus, with a swap between the third and the last sentence.

- CLASS: the class to be predicted (1 if the target follows the prompt, 0 otherwise).

For sub-task 2, we mixed data from the two sources and released a single dataset with the following structure:

- ID: a simple identifier for the entry;
- TEXT: the 4-sentence prompt to be evaluated;
- MEAN: the coherence score of the text to be predicted, based on the mean of the human judgements collected;

In the context of DisCoTEX, for both sub-tasks participants could leverage further external resources to enhance their models, with the exception of Wikipedia and mTEDx data.

4. Evaluation measures

We defined the following evaluation metrics for each sub-task:

- For sub-task 1: the evaluation metric is Accuracy (the ratio between correctly predicted samples and all processed samples) obtained by each system in the test set. We also reported Precision, Recall and F-score for the two classes;
- For sub-task 2: the evaluation metric is the harmonic mean between Pearson and Spearman correlation coefficients between the participants' scores and test set scores.

Baseline The baseline for both tasks has been computed by employing one-hot vectors representations: For sub-task 1 we extracted the one hot vector for each sentence s_i in the input prompt $P = \{s_1, s_2, ..., s_n\}$, as well as for the target sentence t. The distance between the prompt P and the target sentence t, $\mathscr{D}(P,t)$ is computed as the average distance between each sentence s_i from the prompt, and the target sentence t based on Hamming distance coefficient *Dist*:

$$\mathscr{D}(P,t) = \frac{1}{n}\left(\sum_{i=0}^{n} Dist\langle s_i, t\rangle\right).$$

To decide whether the target sentence t is coherent with the paragraph P we computed the median distance value across the whole training dataset, and we used this as a threshold: all the test samples with a distance value under the median have been considered coherent, incoherent otherwise.

For sub-task 2 we first extracted the one-hot vectors from each sentence s_i in the input prompt $P = \{s_1, s_2, ..., s_n\}$:

$$\overrightarrow{v_1} \leftarrow s_1, \overrightarrow{v_2} \leftarrow s_2, ..., \overrightarrow{v_n} \leftarrow s_n.$$

Then we computed the proximity between each consecutive vectors pair $\langle \overrightarrow{v_i}, \overrightarrow{v_{i+1}} \rangle \in V$ through Jaccard distance metric *Dist*, thereby resulting in (n-1) distance scores, grasping the degree of semantic overlap between each two neighbouring sentences. In order to compute the coherence score for the paragraph P, $\mathscr{C}(P)$, we averaged the scores featuring each pair of adjacent sentences:

$$\mathscr{C}(P) = \frac{1}{n-1}\sum_{i=1}^{n-1} Dist(\overrightarrow{v_i}, \overrightarrow{v_{i+1}}).$$

5. Participants

We received a total of 3 submissions for sub-task 1 and 2 submissions for sub-task 2 from 3 different teams. Each submission had the option to include up to three different runs. The strategies used to approach the task are all very different from each other. Teams that participated in both sub-tasks opted to use the same strategy for both challenges. None of the systems chose to utilize

Team	Members	Affiliations	sub-task 1	sub-task 2	# Runs
MPG	3	Sony Computer Science Laboratories Paris, France Enrico Fermi's Research Center (CREF), Rome, Italy, Sapienza University of Roma, Italy	✓	X	4
IUSSNets	3	Iuss Pavia, Italy	✓	✓	9
ExtremITA	4	Università degli Studi di Roma Tor Vergata, Università di Torino, Italy	✓	✓	6

Table 3
Teams participating in EVALITA 2023 DisCoTeX shared task.

additional resources apart from the official datasets. Further information regarding the task participation can be found in Table 5.

MPG

The MPG team [19] utilized the tree-based classifier LightGBM incorporating a set of explicitly engineered features aiming at comparing the prompt and target with respect to several metrics such as TF-IDF vectors, counts of upper case words, tenses, punctuation, words, and characters, as well as sentence embeddings extracted from Sentence-BERT [20]. They exclusively participated in sub-task 1 with two runs .

IUSSNets

The IUSSNets team [21] employed fine-tuning techniques on four distinct Italian language models: BERT-ita [22], Electra-ita [22], Umberto [23], and Bertino [24] separately for each sub-task. For sub-task 1, they submitted three BERT fine-tuned models: the first fine-tuned on Wikipedia (BERT 1), the second on mTEDx (BERT 2), and the third on both (BERT 3), achieving the second-place score. For sub-task 2, they submitted BERT, Bertino, and Electra fine-tuned models, once again securing the second position, primarily due to the performance of the Electra model.

ExtremITA

The ExtremITA team [25] competed using two multi-task Language Models. The first model (ExtremITA-iT5) is an encoder-decoder based on iT5-small [26], while the second model (ExtremIT-LLaMA) is a decoder based on Camoscio [27], the Italian version of LLAMA [28]. These models largely differ in number of parameters: iT5-small has approximately 110 Million parameters, while the used version of Camoscio has 7 Billion parameters. Both models underwent joint fine-tuning on all EVALITA 2023 tasks and sub-tasks, leveraging prompting techniques. For both DisCoTeX the extremIT5 model received each instance of the dataset preceeded by the task and sub-task name and it produced the predicted label or score as output. Conversely, the extremITLLaMa model, which requires a structured prompt, was provided with a textual description of the task and the desired output format specification. For sub-task 1 the prompt is: *"Le due frasi precedenti, separate da '[SEP]', sono coerenti tra loro? Rispondi sì o no"*; while for sub-task 2 the prompt is: *"Quanto è coerente questa frase in una scala da 0 a 5?"*. Their team emerged as the winner across both DisCoTeX sub-tasks and datasets, thanks to the LLAMA-based model. However, the iT5-based model performed considerably worse, especially in the second sub-task where it remained below the baseline.

Team	Model	Accuracy
EXTREMITA	LLAMA	0.815
IUSSNETS	BERT	0.723
MPG	LGBM	0.595
BASELINE	HAMMING	0.525

Table 4
DisCoTeX leaderboard on sub-task 1.

6. Results

Tables 4 and 5 report the leaderboard of systems taking part in sub-task 1 and sub-task 2, respectively. Note that, for the purpose of the official ranking, for sub-task 1 we considered the accuracy of the best run, and we further computed the mean between the best result/run on Wiki and the best result/run on mTEDx data. Conversely, for sub-task 2 we first computed both Pearson and Spearman correlations, then we applied the harmonic mean between the two measures.

As it can be seen, all systems outperform the baseline in both sub-tasks. The best performance was achieved by the team extremITA with the system based on the LLAMA model.

Team	Model	r	ρ	HM
EXTREMITA	LLAMA	0.66	0.65	0.65
IUSSNETS	ELECTRA	0.65	0.62	0.63
IUSSNETS	BERT	0.64	0.60	0.62
IUSSNETS	BERTINO	0.50	0.48	0.49
BASELINE	JACCARD	0.10	0.13	0.11
EXTREMITA	T5	0.06	0.06	0.06

Table 5
DISCOTEX leaderboard on sub-task 2. Reported figures express Pearson (r) and Spearman (ρ) correlations together with their harmonic mean (HM).

7. Analysis and Discussion

To better examine the influence of genre on the automatic modeling of text coherence, we evaluated all submitted systems considering the two datasets separately. The outcomes for sub-task 1 are presented in Table 6, which displays the accuracy scores for both the positive and negative classes of the mTEDx and WIKI data, along with precision, recall, and F1 measures. From this fine-grained perspective, it appears that sub-task 1 is more effectively tackled using the dataset extracted from Wikipedia, with an average accuracy of 0.71, compared to the subset extracted from mTEDx talks, which achieved an average accuracy of 0.63. This discrepancy in performance could be attributed to the structural differences in the texts. Language models indeed are more frequently exposed to the encyclopedic language found in Wikipedia, whereas the lecture-style transcriptions of spoken language in mTEDx talks pose a greater challenge. Furthermore, determining whether the target sentence directly follows the prompt may be more difficult in the transcription of a mTEDx talk. Indeed, unlike traditional lectures, mTEDx talks are a type of public speech that is designed to engage a broad audience by exploring a single idea with unwavering focus. Focusing on the single submissions, the extremITA team outperforms the other participants on both datasets with the run corresponding to the LLaMA based model. Such improvement over competitors may be due to the large amount of model parameters (7B) together with the multi-task setting. The results obtained by the IUSSNets team are comparable to those obtained by the first run of extremITA: although both models are based on Transformers architecture, the results obtained by the first and second runs on mTEDx and Wiki by the IUSSNets team seem to be more stable in terms of Precision and Recall. Regarding the MPG team's results, the higher difference from other competing models may indicate the lack of discourse-level features, these could be beneficial to fully grasp the properties of the word sequence.

The detailed results for the sub-task 2 are reported in Table 7. In contrast to the first sub-task, in the second one

Team	Model (run)	Ted						
		0			1			Accuracy
		P	R	F1	P	R	F1	
IUSSNETS	BERT(1)	0,71	0,70	0,70	0,70	0,71	0,71	**0,70**
	BERT(2)	—	—	—	—	—	—	—
	BERT(3)	0,50	0,28	0,36	0,50	0,71	0,59	0,50
MPG	LGBM(1)	0,54	0,78	0,64	0,60	0,32	0,42	0,55
	LGBM(2)	0,55	0,67	0,60	0,57	0,45	0,50	**0,56**
EXTREMITA	T5	0,77	0,57	0,66	0,66	0,83	0,74	0,70
	LLAMA	0,78	0,80	0,79	0,79	0,78	0,79	**0,79**
BASELINE	HAMMING	0,51	0,43	0,47	0,51	0,59	0,54	0,51

Team	Run ID	Wiki						
		0			1			Accuracy
		P	R	F1	P	R	F1	
IUSSNETS	BERT(1)	—	—	—	—	—	—	—
	BERT(2)	0,75	0,71	0,73	0,72	0,76	0,74	**0,74**
	BERT(3)	—	—	—	—	—	—	—
MPG	LGBM(1)	0,61	0,69	0,65	0,64	0,56	0,60	0,62
	LGBM(2)	0,63	0,64	0,63	0,63	0,62	0,62	**0,63**
EXTREMITA	T5	0,85	0,49	0,62	0,64	0,91	0,75	0,70
	LLAMA	0,87	0,81	0,84	0,82	0,88	0,85	**0,84**
BASELINE	HAMMING	0,54	0,50	0,52	0,54	0,58	0,56	0,54

Table 6
Detailed results on sub-task 1 considering mTEDx (top) and Wikipedia (bottom) datasets.

the best results are generally obtained on the subset of mTEDx talks. Indeed, the system submitted by IUSSNets seems to suffer particularly from the Wiki subset: the third run of IUSSNets has a difference of 0.24 and 0.27 in terms of Pearson and Spearman correlations respectively, while the second run of extremITA only 0.12 and 0.11 respectively. The unexpected best performance of systems on mTEDx talks warrants further investigation, especially considering the higher variance in coherence scores obtained for prompts extracted from TED talks (std: 0.92) than those extracted from Wiki data (std: 0.63).

Team	Model (run)	Ted		Wiki		Pert.		Non-Pert.	
		r	ρ	r	ρ	r	ρ	r	ρ
EXTREMITA	T5	0,12	0,13	0,03	0,01	0,06	0,07	0,08	0,04
	LLAMA	0,65	0,66	**0,53**	**0,55**	**0,65**	0,66	**0,66**	**0,6**
IUSSNETS	BERTINO	0,46	0,48	0,31	0,3	0,53	0,55	0,48	0,35
	BERT	0,66	0,68	0,44	0,44	0,63	0,64	0,66	0,51
	ELECTRA	**0,69**	**0,71**	0,45	0,44	**0,65**	**0,69**	**0,66**	0,49
BASELINE	JACCARD	0,12	0,18	0,13	0,16	0,08	0,14	0,11	0,09

Table 7
Detailed results on sub-task 2, in terms of Pearson (r) and Spearman (ρ) correlations for both subsections. We additionally reported correlation scores for perturbed (Pert.) data and unperturbed data (Non-Pert.).

8. Conclusion

We presented the results of the DISCOTEX task, held within EVALITA 2023 [14]. The task was divided in two sub-tasks: the first one challenged participants to propose systems able to discriminate between coherent and incoherent textual passages, where the latter have been artificially created to gradually undermine local coher-

ence within text. The second one intended to model the human perception of text coherence by predicting the average score attributed to human raters to a text. A novel dataset was developed for this task comprising texts from two different domains, representative of a written and spoken language variety in order to investigate the role of modality on the automatic modeling of coherence. Three teams participated in the task and submitted a total of 19 runs. Notably, the ExtremITA team secured the first position in both sub-tasks with their system based on the largest decoder model proposed. However, it is worth highlighting that smaller models with fewer parameters also demonstrated comparable performance, indicating their effectiveness in capturing discourse-related information. Quite surprisingly, the results of sub-task 2 revealed that systems were more proficient in predicting coherence scores for TEDx talks compared to Wikipedia texts, which calls for further investigation by also expanding the current dataset of human evaluated texts. Future plans involve extending the DisCoTeX task to a multilingual perspective, enabling coherence modeling exploration across different languages using reproducible data collection processes in languages with available Wiki and TED resources.

Acknowledgements

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] T. A. Van Dijk, W. Kintsch, Strategies of discourse comprehension, Academic Press, New York, 1983.

[2] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (1995) 203–225. URL: https://aclanthology.org/J95-2003.

[3] R. Barzilay, M. Lapata, Modeling Local Coherence: An Entity-Based Approach, Computational Linguistics 34 (2008) 1–34. URL: https://doi.org/10.1162/coli.2008.34.1.1. doi:10.1162/coli.2008.34.1.1. arXiv:https://direct.mit.edu/coli/article-pdf/34/1/1/1798481/coli.2008.34.1.1.pdf.

[4] M. Elsner, E. Charniak, Disentangling chat with local coherence models, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1179–1189. URL: https://aclanthology.org/P11-1118.

[5] D. Tien 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), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1320–1330. URL: https://aclanthology.org/P17-1121. doi:10.18653/v1/P17-1121.

[6] J. Li, D. Jurafsky, Neural net models of open-domain discourse coherence, ArXiv abs/1606.01545 (2017).

[7] A. Lai, J. R. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, CoRR abs/1805.04993 (2018). URL: http://arxiv.org/abs/1805.04993. arXiv:1805.04993.

[8] 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.

[9] 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.

[10] 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.

[11] 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. URL: https://doi.org/10.1162/tacl_a_00388. doi:10.1162/tacl_a_00388.

[12] M. Chen, Z. Chu, K. Gimpel, Evaluation benchmarks and learning criteria for discourse-aware sentence representations, arXiv preprint arXiv:1909.00142 (2019).

[13] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, in: Proceedings of the First Workshop on Computational Approaches to Discourse, Association for Computational Linguistics, Online, 2020, pp. 102–112. URL: https://aclanthology.org/2020.codi-1.11. doi:10.18653/v1/2020.codi-1.11.

[14] 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.

[15] D. Brunato, C. Chesi, F. Dell'Orletta, S. Montemagni, G. Venturi, R. Zamparelli, AcCompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).

[16] L. Gregori, M. Montefinese, D. P. Radicioni, A. A.

Ravelli, R. Varvara, Concretext @ EVALITA2020: The Concreteness in Context Task., EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).
[17] G. Caliendo, The popularisation of science in web-based genres, The language of popularisation: Theoretical and descriptive models 3 (2012) 101–132.
[18] D. Biber, S. Conrad, R. Reppen, Corpus linguistics: investigating language structure and use., Cambridge University Press, Cambridge, 1998.
[19] M. Galletti, P. Gravino, G. Prevedello, MPG at DisCoTex: Predicting text coherence by tree-based modelling of linguistic features, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[20] 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.
[21] E. Zanoli, M. Barbini, C. Chesi, IussNets at DisCoTex: A fine-tuned approach to coherence, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[22] S. Schweter, Italian BERT and ELECTRA models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.
[23] L. Parisi, S. Francia, P. Magnani, UmBERTo: an Italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.
[24] M. Muffo, E. Bertino, BERTino: an Italian DistilBERT model, Computational Linguistics CLiC-it 2020 (2020) 317.
[25] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA2023: Multi-task sustainable scaling to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[26] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for Italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.
[27] A. Santilli, Camoscio: An Italian instruction-tuned LLaMa, https://github.com/teelinsan/camoscio, 2023.
[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMa: Open and efficient foundation language models, 2023. arXiv:2302.13971.