MPG at DisCoTeX: Predicting Text Coherence by Tree-based Modelling of Linguistic Features

Martina Galletti[1,2,3], Pietro Gravino[1,2] and Giulio Prevedello[1,2]

[1]*Sony Computer Science Laboratories Paris, 6, Rue Amyot, 75005, Paris, France*

[2]*Enrico Fermi's Research Center (CREF), via Panisperna 89A, 00184, Rome, Italy*

[3]*Department of Computer, Control and Management Engineering (DIAG) "Antonio Ruberti", Sapienza University of Rome, via Ariosto 25, Rome, 00185, Italy*

Abstract

Automatic text coherence modelling plays a crucial role in natural language processing tasks, such as machine translation, summarisation, and question answering. Moreover, text coherence is fundamental to reading comprehension and readers' engagement, essential to a number of application domains. In this report, we report progress for the Assessing Discourse Coherence in Italian Texts task from EVALITA-23, whose goal is to address automatic coherence detection. The task was challenged by extracting linguistic features used to train a machine learning classifier, leading to minor improvement over the baseline. The feature importance analysis revealed semantic features' relevance, providing indications for future feature engineering and modelling efforts.

Keywords

Natural language processing, text coherence, Italian language, machine learning

1. Introduction

Coherence is an essential quality to facilitate comprehension and readers' engagement. Text coherence measures the quality of organization in the structure of a text and the extent to which a reader can follow the relationships between sentences and paragraphs. Several text coherence models exist in the literature which aim to distinguish between coherent and incoherent documents. This distinction significantly affects some downstream tasks, such as document summarisation, automatic essay assessment, and the creation of natural language [1, 2, 3]. Traditional approaches often focus on linguistic features such as discourse markers [4, 5], textual connectives [6], entity-grid based approaches [7, 8] which takes inspiration from Centering Theory [9] to capture coherence [7, 10]. Recent approaches rely on neural architectures and use Convolutional Neural Networks (CNNs) over an entity-based representation of text [11, 12], Sequence to Sequence Models [3, 13] and Multi-Task Learning [14]. Nevertheless, language models still face challenges in capturing and predicting global coherence across longer texts, and targeted evaluation paradigms are still being implemented [15, 16, 17, 18, 19].

In addition to being fundamental in a number of NLP tasks, the automatic assessment of coherence in a text could have significant applications in hybrid speech and language therapy. Coherence evaluation can assess various text sources, such as educational materials, therapy resources, or online content, commonly used in therapy sessions. By utilizing automatic metrics, therapists can have measures to judge the complexity and accessibility of these texts faster, selecting appropriate materials that align with their patients' reading abilities. This ensures therapy resources are tailored to individual needs, promoting engagement, comprehension, and progress.

In the context of news media, coherence measurement is important to evaluate news items' readability but is also involved in the assessment of the alignment between an article's content and headline. Sensationalized or exaggerated headlines are sometimes used to attract attention and gain clicks to online news, a practice termed clickbaiting that undermines trust in news media. The automatic detection of text coherence can empower readers to focus on accurate journalistic products, thus contributing to the transparency and trust in news media and encouraging responsible reporting of information.

In this work, text coherence is addressed by evaluating whether a target sentence follows a small prompt paragraph (made of a few sentences) in a coherent (or not) way. This problem is a subtask from the challenge "Assessing Discourse Coherence in Italian Texts" (DisCoTeX) [20] presented at EVALITA 2023 [21]. In this report, the feature extraction procedure and the modelling strategy are described in Section 2, while the results are illustrated in Section 3 and discussed in Section 4.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, September 07–08, 2023, Parma, Italy

martina.galletti@sony.com (M. Galletti); pietro.gravino@sony.com (P. Gravino); giulio.prevedello@sony.com (G. Prevedello)

0000-0002-0937-8830 (P. Gravino); 0000-0002-9857-2351 (G. Prevedello)

2. Description of the system

The data provided were cleaned and pre-processed in a series of systematic steps. After removing irrelevant characters, the dataset provided by the task organisers was normalised in a standardised format, i.e. lemmas, to break the text into meaningful units. Then, words with less than three characters were removed, keeping stop words. This was done because short words often carry less semantic meaning than longer words, and thus they could increase the vocabulary size of a model without contributing significantly to its training. By removing these words, some potential noise in the training could be removed. On the other hand, stop words, especially conjunctions, were kept for their role in connecting sentences across a text and preserving syntactic and grammatical relationships during the training phase, even if they do not carry a semantic meaning *per se*. After lemmatization, we added more info to the provided data by computing the length of words and sentences. This was done because the length of the words and/or sentences can provide insights into the complexity and readability of the text, which can, in turn, impact its intrinsic coherence. Longer words can indicate, in fact, a more complex and domain-specific vocabulary, while longer sentences could indicate the presence of multiple sub-clauses, which could impact the overall coherence. Moreover, an abrupt variation in the length of words and sentences could indicate an unbalance in the structure of the text, thus endangering its linguistic coherence. For similar reasons, the statistics of the uses of the different tenses in sentences were also computed since the usage of appropriate tenses ensures temporal consistency and logical progression of information. Afterwards, we extracted lexical features such as word frequency, for which we used document term matrix, Term Frequency - Inverse Document Frequency (TF-IDF), and sentence embeddings, i.e. Sentence-BERT (SBERT) [22].

To compress high-dimensional vectors, from the TF-IDF analysis and the sentence embedding, Uniform Manifold Approximation and Projection (UMAP) was used to reduce their dimension down to 30 components [23]. UMAP was chosen over the principal component analysis method as it tends to preserve local distances better. Meanwhile, compared to the t-distributed stochastic neighbour embedding method, it is faster and better preserves the global data structure.

Finally, prompt and target were compared by means of the statistics mentioned above, resulting in the following list of features for each data point:

- *tfidf_raw_wjac*: weighted Jaccard distance between prompt's and target's TF-IDF vectors;
- *tfidf_raw_cos*: cosine distance between prompt's and target's TF-IDF vectors;
- *tfidf_red_euc*: euclidean distance between prompt's and target's TF-IDF vectors projected by UMAP;
- *tfidf_red_cos*: cosine distance between prompt's and target's TF-IDF vectors projected by UMAP;
- *upper_case_word_count_t_over_tp*: the number of upper case words in target divided by their sum in prompt and target;
- *word_density_t_over_tp*: word density in target divided by the sum in prompt and target;
- *punctuation_count_t_over_tp*: the number of punctuation marks in target divided by their sum in prompt and target;
- *char_count_t_over_tp*: the number of characters in target divided by their sum in prompt and target;
- *word_count_t_over_tp*: the number of words in target divided by their sum in prompt and target;
- *tense_tar_per*: the size of the set of tenses in both target and prompt divided by the one in the target only;
- *tense_iou*: the size of the set of tenses in both target and prompt divided by the one in either target or prompt;
- *ent_tar_per*: the size of the set of entities in both target and prompt divided by the one in the target only;
- *ent_iou*: the size of the set of entities in both target and prompt divided by the one in either target or prompt;
- *sbert_umap1_pro*: first component of the 2d UMAP projection of the average vector from the embedding of prompt's sentences;
- *sbert_umap2_pro*: second component of the 2d UMAP projection of the average vector from the embedding of prompt's sentences;
- *sbert_umap1_tar*: first component of the 2d UMAP projection of the vector from the embedding of target's sentence;
- *sbert_umap2_tar*: second component of the 2d UMAP projection of the vector from the embedding of target's sentence;
- *sbert_red_euc*: euclidean distance between 30d UMAP projection of the average vector from the embedding of prompt's sentences and the vector from the embedding of target's sentence;
- *sbert_red_cos*: cosine distance between 30d UMAP projection of the average vector from the embedding of prompt's sentences and the vector from the embedding of target's sentence;
- *emb_cos_mean*: average of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's;
- *emb_cos_max*: maximum of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's;

- *emb_cos_last*: the pairwise cosine distances between the vector embedding of the prompt's last sentence and the target's;
- *emb_cos_min*: minimum of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's.

These features were then passed to a machine learning model that classifies whether the target sentences were coherently following the prompt text, thus tackling the Subtask 1 of the challenge.

The classifier model of choice was LGBMClassifier, a popular machine learning solution that combines computational efficiency with good predictive performances in various problems. The model was imported from the LightGBM gradient boosting framework that uses tree-based learning algorithms [24], and the binary cross-entropy was set as the objective function for the training. Model's hyperparameters were selected by stratified 10-fold cross-validation on shuffled data [25], exhaustively searching the space of hyperparameters ($\mathit{num\_leaves} \in \{24, 25, 26\}$, $\mathit{max\_depth} \in \{5, 6, -1\}$, $\mathit{learning\_rate} \in \{0.009, 0.01, 0.011\}$, $\mathit{n\_estimators} \in \{185, 190, 195\}$) for the combination with best overall accuracy. This search space of hyperparameters was defined empirically, starting with intervals centred around extreme parameters (*num_leaves* 50, *max_depth* 10, *learning_rate* 0.01, and *n_estimators* 1000). Then, for every new training instance, those intervals were re-centred at the previous best-fitting value if such value stood at the extremity of the interval. Otherwise, the interval was made more narrow. The random seed was set to 42 for reproducibility.

The model's performance was evaluated against the baseline provided by the challenge organizers. See [20] for more details.

3. Results

Training the system on the data available for Subtask 1, the model achieved an accuracy of 0.595 on the test set, improving upon the challenge baseline (0.525) by only 0.07 points. To provide some insights into these performances, the confusion matrix on the training data and the importance of the features are shown in Figure 1 and Figure 2, respectively. Finally, the relevant hyperparameters, resulting from the grid-search cross-validation, are reported in Table 1.

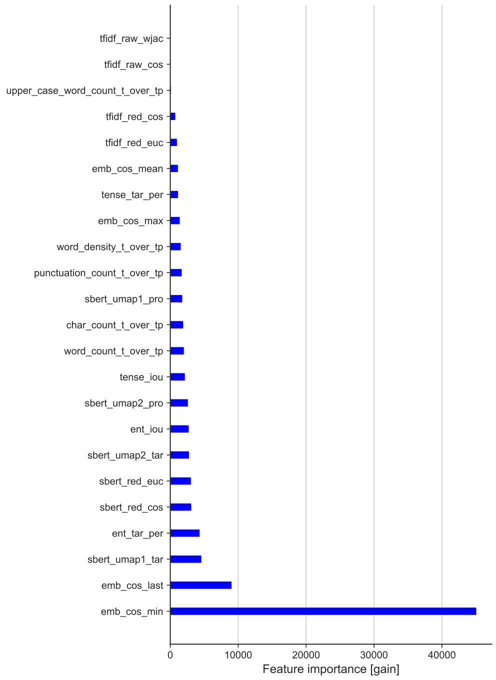

Figure 1: Importance of features from the decision tree-based classifier model.

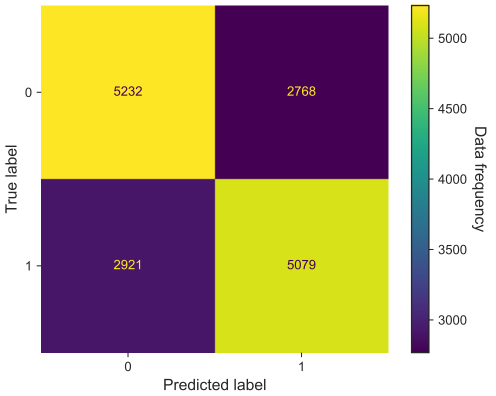

Figure 2: Confusion matrix from the classifier model's predictions on the training data.

4. Discussion

The features' importance highlights that standard frequency statistics (such as comparisons between prompt and target on TF-IDF vectors and counts of upper case words, tenses, punctuation, words, and characters) are not very informative in predicting coherence for the system employed. Yet "*ent_tar_per*", the proportion of target's entities also present in the prompt's text, was rather important. Semantic information seemed the most relevant, as supported by the many important features extracted by leveraging the sentence embedding model. Of note, "*sbert_umap1_tar*", the projection of the target's sentence embedding, was quite important although not derived from the comparison between prompt and target.

Table 1
Best-fitting hyperparameters from the grid-search cross-validation of the classifier model.

Hyperparameter	Value
boosting_type	gbdt
colsample_bytree	1
importance_type	gain
learning_rate	0.01
max_depth	6
min_child_samples	20
min_child_weight	0.001
min_split_gain	0
n_estimators	185
num_leaves	25
objective	binary
random_state	42
metric	accuracy

Finally, the most important features resulted from the aggregation of the pairwise cosine distances between the sentence embedding of the target sentence against those from the prompt sentences. While the average and the maximum of these distances seemed not much important (“*emb_cos_mean*'', “*emb_cos_max*''), the minimum distance and the distance between the target's and prompt's last sentences (“*emb_cos_min*'', “*emb_cos_last*'') were the two most important features. These findings suggest that coherence is elicited by one or a few proximal sentences, while the rest might be of secondary importance.

Our results suggest that including syntactic and discourse-level features might lead to improved performances. Syntactic features, such as part of speech tagging, dependency relationships, or parsing trees, can provide insights into sentence structure and overall grammatical coherence. Moreover, discourse-level features, such as entity co-reference, readability metrics, argumentative structure, discourse markers or topics progression-related features, could assess the flow of ideas in the documents provided. Future work will address the extraction of discourse structure and syntactic features to enable our model to assess a certain text's logical connections, organization and grammatical structure.

The moderate performance improvement might also suggest limitations of standard machine learning models. This could be due to several reasons. First, these models have limited representation capacity of semantic relationships compared to deep learning models, as they lack the sequential modelling needed to represent a text's underlying coherence. Moreover, they lack automatic contextual understanding, focusing more on provided features that might not capture the global context appropriately. Finally, they generally struggle with high-dimensional data. If the number of features is high, as in text coherence, the model could perform sub-optimally, with increased computational complexity. Future improvements on the modelling will include the use of deep learning techniques such as CNNs [11, 12] and Sequence Models [3, 13].

Acknowledgments

This work has been supported by the Horizon Europe VALAWAI project (grant agreement number 101070930). We also wish to thank the Evalita-23 organizers for organizing the task and emphasizing the significance of text coherence measures for the Italian language. The creation and annotation of the corpus have been instrumental in advancing the field of Natural Language Processing for the Italian language and fostering community interest in coherence assessment. Your efforts will enable us to develop AI-driven methods for fostering comprehension assessment applied to both the infosphere and hybrid speech and language practices.

References

[1] Y. Jernite, S. R. Bowman, D. Sontag, Discourse-based objectives for fast unsupervised sentence representation learning, arXiv preprint arXiv:1705.00557 (2017).

[2] Y. Wu, B. Hu, Learning to extract coherent summary via deep reinforcement learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[3] J. Li, D. Jurafsky, Neural net models of open-domain discourse coherence, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 198–209. URL: https://aclanthology.org/D17-1019. doi:10.18653/v1/D17-1019.

[4] 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.

[5] M. Zhang, V. W. Feng, B. Qin, G. Hirst, T. Liu, J. Huang, Encoding world knowledge in the evaluation of local coherence, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1087–1096.

[6] G. Albertin, A. Miaschi, D. Brunato, On the role of textual connectives in sentence comprehension: A new dataset for italian., in: CLiC-it, 2021.

[7] R. Barzilay, M. Lapata, Modeling local coherence:

An entity-based approach, Computational Linguistics 34 (2008) 1–34.

[8] M. Elsner, E. Charniak, Extending the entity grid with entity-specific features, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 125–129.

[9] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modelling the local coherence of discourse, IRCS Technical Reports Series (1995).

[10] M. Lapata, R. Barzilay, et al., Automatic evaluation of text coherence: Models and representations, in: Ijcai, volume 5, 2005, pp. 1085–1090.

[11] 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.

[12] H. C. Moon, T. Mohiuddin, S. Joty, X. Chi, A unified neural coherence model, arXiv preprint arXiv:1909.00349 (2019).

[13] 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4328–4339. URL: https://aclanthology.org/D18-1464. doi:10.18653/v1/D18-1464.

[14] Y. Farag, H. Yannakoudakis, Multi-task learning for coherence modeling, arXiv preprint arXiv:1907.02427 (2019).

[15] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, arXiv preprint arXiv:2105.03495 (2021).

[16] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, arXiv preprint arXiv:2011.06306 (2020).

[17] A. Lai, J. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, arXiv preprint arXiv:1805.04993 (2018).

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[23] L. McInnes, J. Healy, N. Saul, L. Grossberger, Umap: Uniform manifold approximation and projection, The Journal of Open Source Software 3 (2018) 861.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.