National Research University Higher School of Economics

as a manuscript

Elizaveta Goncharova

INFORMATIVE DISCOURSE FEATURE SELECTION FOR ANALYSIS OF TEXTUAL DATA

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow-2022

The PhD dissertation was prepared at National Research University Higher School of Economics

Academic Supervisor: Sergei O. Kuznetsov, Doctor of Science, Professor of National Research University Higher School of Economics

DISSERTATION TOPIC

Dissertation relevance. Throughout recent decades, natural language processing (NLP) tasks have uncovered many sophisticated problems in the machine learning (ML) domain, leading to the rapid development of NLP techniques. The object of research in the NLP domain is a spoken or, more commonly, written text in natural language. In contrast, the research objectives vary depending on the specific NLP tasks: text classification, text generation, question answering, summarization, and many others.

Generally, NLP models aim to construct informative latent representations of the raw textual data. These latent representations should encapsulate all the relevant linguistic information about a text that is informative for further downstream tasks such as classification or generation. Conventional NLP techniques have been focused on generating manually constructed linguistic features for text representation. This feature-based approach is resource-intensive and time-consuming as it requires extensive expertise from well-qualified specialists. The manually constructed features are task-specific, making each novel task or domain require customization of the constructed feature subset. Hence, these algorithms are not flexible with respect to the new tasks. Within the development of deep learning (DL) techniques, the manually constructed features have been replaced with the automatically learned representations of the raw texts. These textual representations can be extracted from the DL models trained on extensive collections of text data, and they are more general and flexible than the manual features.

Deep learning models are the typical approach for representation construction, starting with the simple feed-forward neural networks that allow mapping words into an informative low-dimensional vector representation [37, 40], followed by the recurrent neural networks (RNN) [8], and, finally, the transformer-based models: BERT [12], or series of GPTs [43], etc. Transformer-based models are pre-trained on the large text corpora to accomplish some general language tasks; they serve as a powerful and convenient technique for constructing the text representation that can be further leveraged for many downstream tasks [44, 36] without the need for further training.

Despite their great success, scholars have shown [38, 60] that, when following a data-driven paradigm, big pre-trained models can have a low quality

of linguistic knowledge encoding, degrading the performance of these models for some NLP problems. Another critical obstacle for most pre-trained models was outlined in [24, 56, 20], which discuss the inability of the models to process lengthy text passages. Probing experiments have shown that pre-trained models fail to capture long-range dependencies between the text spans [51, 50], creating illogical outputs for long and detail-riddled input sequences. While these drawbacks can be overcome for some tasks, several use cases are highly dependent on the text's linguistic organization and suffer from insufficient vector representation provided by these models. Several problematic tasks are machine reading comprehension for lengthy passages, argumentation mining, and fake news detection. In general, the domain helps determine when the system should observe the complicated structure of the documents rather than just the words themselves.

In such condition, a central issue in the modern NLP domain is not only to increase the size of the models and training corpora, but also to create some approaches models enabling the pre-trained models to leverage relevant information about the language in order to provide more accurate and consistent solutions to the performed NLP problems.

This dissertation addresses this issue and considers how we could explicitly incorporate the discourse structure of a text into transformer-based models without significant modification required for the model's architecture. We introduce two novel techniques for incorporating discourse structure into the BERT model. The former is associated with a discourse-aware MLM task used for pre-training purposes. The latter is a discourse-aware self-attention mechanism that can provide additional navigation for the model to solve discourse-enhanced tasks. The introduced methods for discourse structure incorporation have been implemented into the BERT model architecture and leveraged to teach this model to encode a paragraph of text with respect to its discourse structure. Discourse-aware models were investigated using two NLP tasks: argumentation classification (AC) and machine reading comprehension (MRC), which can benefit from discourse structure awareness. The results obtained for these tasks have shown that discourse-aware models perform better than their standard variations.

Additionally, this research investigates the use of discourse structure to improve the transformer-based model's explainability. The pre-trained models are hardly interpretable and operate as the *black boxes*, taking the raw texts as

input and providing the output without explaining how the decision is produced. The most common approaches that are leveraged to interpret the model's work are either attention-based, where the attention weights are used as the scoring components defining the importance of the corresponding inputs, or gradient-based [46, 22], where the tokens are analyzed based on their influence on gradient flow changing. While these techniques can provide rational explanations for the model's decision, they are still far from the ideal transparency and interpretability of current representation learning and neural network models. Since the existing approaches are hard to train, they need some improvement. Here, a mixture of the DL-based techniques and the explainable machine learning methods will be investigated to determine if they can improve the model's explainability and interpretability.

Related work. Aside from internal information derived directly from the words and the context, much external information could be injected into the language models to construct more coherent learned representations. External knowledge incorporation has received wide attention for the studies in [32, 50, 6]. The authors tend to analyze whether they can align automatically constructed embeddings with the manually constructed linguistic features that the linguists can initialize as relevant for some specific NLP task.

Extensive research has been done on how additional linguistic information could be encapsulated into the DL models. Typically, the existing techniques aggregate a pair of DL components, where a classical transformer encoder is used to calculate the contextual representation of a text, and the linguistic information is encoded by an additional neural network. For example, it has been proposed that a syntax-guided self-attention mechanism is added at the top of the BERT encoder to enrich its outcome with the syntactic organization of the text [59]; SemBERT [58] allows the BERT to absorb contextual semantics providing improvement of the model for NER task. Despite syntactic and semantic information, the discourse organization of a text is also relevant for LMs. Discourse structure provides an overall logical organization of thoughts expressed by an author [17], i.e., except for a plain text, the model observes main and dependent clauses and the logical connection existing between them. This information could be valuable for some complex NLP tasks that require a model to capture all the hidden dependencies existing in a text and to pay attention only to the relevant words while solving a task.

Goals and objectives of the dissertation.

The **goal** of this research is to assess the relevance of discourse features on the transformer-based models performing a range of NLP tasks where discourse information is relevant and develop approaches to incorporate discourse structure into the transformer-based models implicitly.

To achieve this goal, we established the following **tasks**:

1. Explore the role of discourse analysis for the AC and MRC tasks and demonstrate the insufficiency of the discourse-free models by experimental evaluation of the existing benchmarks.
2. Develop the algorithm to label texts with their discourse structure that the pre-trained transformer-based models can leverage.
3. Introduce a novel approach to incorporate discourse information into the BERT model to enable it to calculate discourse-aware vector representation of a text and check its performance on the AC task.
4. Introduce a novel model for MRC task that ensures discourse-aware attention mechanism.
5. Design an explainability pipeline to retrieve the *rationales* from a text, based on the relevant discourse features.

KEY RESULTS

Scientific novelty. The novelty of the research refers to the following elements:

1. A novel modified MLM task that enables the pre-trained BERT model to capture the discourse structure of a text is introduced.
2. The discourse-aware attention mechanism is first designed to enrich the constructed text representations with discourse information.
3. The explainability pipeline that constructs the grammatically-consistent rationales explaining a model's decision based on the discourse features analysis is developed.

Practical value. In summary, this dissertation has made the following practical contributions.

1. Two novel methods for discourse enrichment of the transformer-based models are presented; the fine-tuned models can be applied for the downstream NLP tasks.

2. A new explainability pipeline is presented that ensures constructing grammatically-consistent rationales that explain the model's decision.
3. The disBERT model was applied to filtering users' reviews in the chatbot for the e-commerce domain [34].
4. The source code of the implemented models and the labeled datasets are available via GitHub[1].

Methodology and research methods. The research involved the application of Machine learning, including Deep learning methods, knowledge of probability theory and statistics, discourse analysis, and Formal concept analysis (FCA). A program code is implemented in Python using the pytorch framework for the neural networks implementation and transformers library from the Hugging Face, the numpy and sklearn libraries are also used. The experiments satisfy the principles of reproducibility of the results.

Key aspects to be defended:

1. For the first time the discourse structure is injected into the transformer-based model via segmentation embedding layer of the BERT model.
2. For the first time the discourse-aware self attention mechanism is applied to inject linguistic knowledge into the model.
3. The algorithm for converting the discourse tree into the dependency discourse graph that allows to encode only the relevant features is proposed.
4. The novel research has been conducted introducing the local explainability pipeline that uses discourse information to provide the interpretable rationales.

Author's contribution.

The discursive-aware disBERT model and the modified masked language modeling task were proposed and implemented by the author of the dissertation. An algorithm for converting a discursive parse tree into a discursive dependency graph, taking into account relevant features, is also the result of the author's personal work. The BERT model, enriched with the discourse-aware attention mechanism was introduced by the author of the dissertation in collaboration with the co-authors Boris Galitsky and Dmitry Ilvovsky. The author of the disserta-

[1]https://github.com/lizagonch/Discourse-BERT

tion proposed to use information from the discourse dependency graph for the attention mechanism, and experiments to obtain vector representations for each of the elementary discourse units were conducted by the author. The author of the dissertation also collected and labeled the datasets leveraged in the experiments. The method for constructing *the rationales* is the result of the author's research, while the algorithm used to find relevant discursive schemes using the FCA method was developed in collaboration with the academic supervisor Sergei Kuznetsov.

PUBLICATIONS AND APPROBATION OF RESEARCH

The main results reported in this research were presented on the International conferences and workshops and published in the sources indexed by Scopus and Web of Science.

Second-tier publications:

1. Goncharova E.[1] Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension / Galitsky B., Ilvovsky D., Goncharova E. // Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 444-453, 2021. (Scopus)
2. Goncharova E. Relying on Discourse Trees to Extract Medical Ontologies from Text / Galitsky B., Ilvovsky D., Goncharova E. // Lecture Notes in Computer Science, vol. 12948 LNAI, pp. 215 – 231, 2021. (Scopus – Q2)
3. Goncharova E.[1] Concept-based chatbot for interactive query refinement in product search / Goncharova, E., Ilvovsky, D., Galitsky, B. // CEUR Workshop Proceedings, vol. 2972, pp. 51 – 58, 2021. (Scopus)
4. Goncharova, E. On a chatbot conducting dialogue-in-dialogue / Galitsky, B., Ilvovsky, D., Goncharova, E. // SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference, pp. 118 – 121, 2019. (Scopus, Web of Science)
5. Goncharova E. On a Chatbot Providing Virtual Dialogues / Galitsky B., Ilvovsky D., Goncharova E. // International Conference Recent Advances in Natural Language Processing, RANLP, vol. 2019-September, pp. 382 – 387, Код 155296, 2019. (Scopus)

[1]The author of the dissertation is the main author.

6. Goncharova E.[1] Increasing the efficiency of packet classifiers with closed descriptions / Goncharova E.F., Kuznetsov S.O. // CEUR Workshop Proceedings, vol. 2529, pp. 75 – 88, 2019. (Scopus)

Reports at conferences and seminars:

1. International Conference Recent Advances in Natural Language Processing (RANLP 2021), regular paper: *Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension*, September 1-3, 2021, Varna, Bulgaria (held online).
2. COLING Workshop on Natural Language Processing in E-Commerce (EComNLP 2020), full paper: *On a Chatbot Navigating a User through a Concept-Based Knowledge Model*, December 12, 2020, Barcelona, Spain (held online).
3. 18th Russian Conference on Artificial Intelligence (RCAI-2020), full paper: *FCA-based Approach for Interactive Query Refinement with IR-chatbots*, October 10-16, 2020, Moscow, Russia.
4. The 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2019), demo paper: *On a Chatbot Conducting Dialogue-in-Dialogue*, September 11-13, 2019, Stockholm, Sweden.
5. The 3rd International Workshop Formal Concept Analysis for Knowledge Discovery (FCA4KD), full paper: *Increasing the efficiency of packet classifiers based on closed descriptions*, June 7, 2019, Moscow, Russia.

CONTENTS

In the **introduction** section, the motivation for the presented dissertation is given followed by the research goals and objective, the list of the publications, and the key results of this research. This section also gives an overview of the previous findings that have already been found by other scholars in the observed research area outlining the research gap that we refer to in this dissertation.

The **first chapter** outlines the development of the language models and their applicability to various NLP tasks. The purpose of the modern LMs is to create a complex representation of text that encodes the significant linguistic information that can provide high results on some downstream NLP task.

The dissertation analyzes the pre-trained transformer-based models, therefore, we consider the language modeling (LM-ing) task as the basic pre-train-

ing procedure. Generally, LM predicts the next word w_N in a sequence based on the left context $[w_1,w_2,...,w_{N-1}]$. Therefore, the goal of probabilistic language modeling is either to calculate the probability of the word sequence as $P(w_1,w_2,...,w_{N-1},w_N)$ or to find the probability of the next word in the sequence as $P(w_N|w_1,w_2,...,w_{N-1})$. When generating a novel word, the LM assesses the probability of all the words presented in the vocabulary.

Masked language modeling (MLM) is a modification of the standard LM-ing task has been proposed and successfully applied to the big LMs, such as BERT model and its variations, pre-training. In MLM, the i^{th} word in the input sequence is hidden, and the model is trained to calculate the probability distribution of this word conditioned on its left and right contexts $P(w_i|w_1,...,w_{i-1},w_{i+1},...,w_N)$, where N is the number of words in the input sequence.

Transformer-based models leverage the LM-ing and MLM in combination with huge textual corpora to learn the relevant information about the language during pre-training. This approach has proven to be effective, as, recently, variations of the transformers model pre-trained in this manner outperform all the existing techniques and achieve state-of-the-art results on various NLP benchmarks.

These DL techniques have totally replaced the standard ML methods in the NLP domain that operated on the manually constructed linguistic features carefully retrieved from the texts with the cooperation of the linguists and data scientists.

While vector representations of texts achieve high performance on the bunch of NLP benchmarks, there are still some problems with them. First, these representations often are not able to capture the linguistic peculiaritites corresponding to the text. So that, the model's textual outputs can be grammatically correct, but illogical (like in text generation), or do not rely on the world knowledge that a human is aware of. Besides, these representations are hardly interpretable by humans.

Finally, the first chapter describes the BERT model that we consider in this research as the baseline in more detail. BERT (Bidirectional Encoder Representations from Transformers) is the encoder-only model consisting of 12 self-attention layers to derive the contextualized word representation. Initially, the

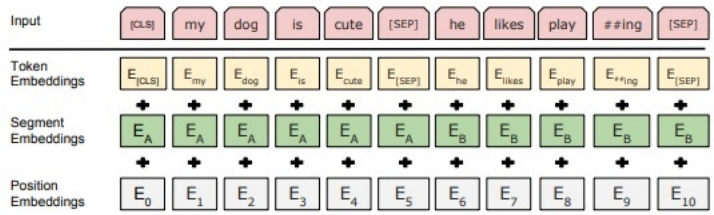

Figure 1 — An input format for the BERT model with the token, segment, and position embeddings.

BERT model was trained using two objectives: MLM and next sentence prediction (NSP). NSP was proposed to pre-train text-pairs representation jointly by predicting whether the second sentence in a pair follows the first one in the original text. The model itself was pre-trained on large corpora of unlabeled texts, namely BooksCorpus (800M words) [140] and English Wikipedia (2,500M words). This variability introduces some architectural modifications to incorporate external linguistic knowledge into the model that we refer to in this research.

In comparison to other transformer-based models that have been trained on one LM-like task, the BERT model is also pre-trained on the NSP task. This training allows the model to successfully perform the natural language inference (NLI) task to identify whether or not one sentence entails another. Due to this task, the model has a special input layer responsible for segmentation encoding that can be used to encode additional information in the input, as described in this research. Figure 1 represents the input layers for the BERT model.

The **second chapter** of the dissertation discusses the question of how well the pre-trained LMs capture linguistic information. In this chapter, we analyze various probing methods introduced in the research designed for checking linguistic awareness of the pre-trained models. Finally, we discuss the obtained results concluding, which type of the linguistic knowledge are captured by the analyzed models, and which of them are under-presented.

The growing interest in pre-trained NLP neural networks has motivated extensive research on discovering the types of linguistic knowledge these models capture. When transformer-based models outperformed all the previous models on NLU benchmarks – such as GLUE [52], SuperGLUE [53] – by a significant margin, the investigations on representation learning took a giant leap. The demonstrated performance gain has aroused considerable interest among scholars who explored whether the transformers can understand structural information about language

or simply provide the distributed language representation based on observing many examples [9].

The linguistic features denote various language components grouped by linguists into five main categories: phonemes, morphemes, lexemes, syntax, and context. The branches of linguistics that study these categories are the following: phonetics, phonology, morphology, syntax, semantics, discourse, and pragmatics.

The probing methods have been presented to explore the learned textual representations and their ability to capture various linguistic characteristics. The former probing techniques were designed to check the model's ability to capture PoS tags [45, 1, 5], morphology [4, 3], or word-sense disambiguation [41]. Tenney et al. [51] extended this research and introduced a general probing architecture that simultaneously covers a wide range of linguistic information, including syntactic, semantic, local, and long-range phenomena. In general, probing tasks are auxiliary tasks performed on the internal representations derived from the pre-trained models. These tasks assess whether these internal representations reflect one or other specific linguistic phenomena either individually or jointly.

Providing wide range of probing tasks, the scholars have come to the conclusion that the multi-head self-attention allows the model to capture some of the linguistic patterns existing in the text via pre-training. Transformer-based models are able to encode the word- and sentence-level linguistic information, however, more complex text-level features, such as discourse structure, is encoded by the models weakly.

Chapter 3 describes the discourse structure and the Rhetorical structure theory (RST) as a framework to describe it that we rely on in this research. This section discusses the existing probing tasks followed by an overview of popular techniques for incorporating discourse into the transformer-based models.

The discourse structure operates at the higher level of text organization. It represents the connections among the units beyond the sentence, up to the entire text [26]. Informally, this structure reflects the development and logical organization of the authors' thoughts throughout the whole text.

This structure can be defined by the existing discourse frameworks, such as RST [35], PDTB [42], or Graph Bank [54]. RST is one of the most popular frameworks, both generally and specifically in this research. Its popularity is due to the availability of the high-quality existing discourse parsers and its ability to

build discourse structure as a combination of semantic and intentional relations that covers a wide range of logical connections existing among the text spans [33].

RST [35] is a discourse relation annotated resource that we rely on in this research for discourse structure retrieval. During discourse parsing, a document can be segmented into non-overlapping text spans – contiguous units for clauses – called elementary discourse units (EDUs). Each EDU can be tagged as either a nucleus or a satellite, where nucleus nodes are more central and satellite nodes more peripheral. Nucleus units consist of the primary information the author expresses in the text, and satellite units contain additional information supporting the one presented in a nucleus. As a result, the discourse structure of the text is represented as a discourse tree (DT). This structure allows relations at the top level to cover relations at the bottom. An example of the DT constructed based on the RST is presented in Figure 3 (b).

There are two main categories of discourse relations: multi-nuclear (N-N) and nucleus-satellite (N-S) relations, also called anti-symmetric relations. The former represents the dependency among the nucleus and satellite units, and the latter holds only between the nucleus units. In anti-symmetric relations, which involve a pair of EDUs, the nuclei are the main components of the relation, and the satellites are their dependents.

The studies have shown that distinction between N-N and N-S relations impacts many NLP tasks, such as anaphora resolution [11, 21] or QA. Researchers noticed that a stronger concentration on different discourse components is needed while answering various questions. For instance, different special *wh-* questions indicate the specific discourse relation and N-S components, e.g., *Attribution* relation is a basis of *what/who is the source* question, where the answer is found in the nucleus, while the satellite in the *Cause* and *Explanation* relations constitute the *why*-questions. This issue was analyzed and implied by us in reference [13, 14, 16].

The discourse-awareness of the pre-trained models have been investigated by the scholars. Koto et al. [29] introduce a unique probing mechanism to check the discourse knowledge incorporation. The authors notice that existing probing techniques focus on identifying syntactic linguistic knowledge. The authors list seven discourse probing tasks that can be utilized by the researchers to investigate, how well DL models understand discourse. The authors conclude that the baseline

BERT model performs well on the NSP tasks, while the results on pure discourse tasks are underwhelming, especially for the early layers. This finding is quite reasonable because BERT was pre-trained using NSP objection.

In our research, we utilize one of the introduced tasks – *discourse connective predictions* – to check the modified model's ability to understand which non-trivial discourse relation connects the observed text spans. This task is also known as discourse markers prediction and identifies discourse markers that connect two elementary discourse units.

Chapter 4 describes the *disBERT* model that has been introduced in the research for incorporation discourse structure into the transformer-based BERT model. This model is a classical BERT model that is further pre-trained on the modified discourse-conditioned MLM task. We show that standard MLM can suffer from insufficiency of the external discourse information about a text, so we propose to modify the existing MLM task by conditioning it on the corresponding discourse relations.

In standard MLM tasks, only the context is required to predict the masked token; however, the contextual information may not be enough to make the correct prediction. Let us consider a simple example that shows that the awarenes of discourse structure can be essential for tokens prediction.

He went out ***while*** *it was not raining. [Condition]*
He went out ***and*** *it was not raining. [Elaboration]*

If the bolded words are masked, the BERT cannot understand which word should replace the mask without additional information about its discourse role in the sentence.

The pre-trained BERT[2] was launched to predict a masked token in the sentence *"He went out [MASK] it was not raining."* There are five most probable predictions made by it, which are *"when," "and,"* a *semicolon*, a *comma*, and *"but."* The pre-trained BERT model decided based on the examples it had seen in the training corpus, and the most probable output was *"when."*

Even though the sentence with *"when"* is grammatically and logically correct, it changes the meaning of a masked sentence. In discourse analysis, such words as *"while," "unless," "when,"* and *"and"* are called *discourse markers*. They

[2]https://demo.allennlp.org/masked-lm

are the special words that could hint at which rhetorical relation connects the text spans. In this case, they could be predicted correctly based on the information about the discourse organization in the text.

The discourse-conditioned MLM task for further pre-training is formalized as follows. We are trying to predict a masked word having information about a context surrounding this word and a label satisfying this part of the input text in the discourse tree. In this case, during the training procedure, the model learns to predict $p\left(\cdot|C_{\backslash w_t},y_t\right)$, where y_t is the label denoting the rhetorical relation, instead of $p\left(\cdot|C_{\backslash w_t}\right)$. This procedure will allow the model to be trained according to the discourse label corresponding to the text span, while other model architecture is kept without modification. This novel discourse-aware MLM objection is as follows:

$$L_{MLM}(D_I|D\backslash\{D_I\}) = \frac{1}{K}\sum_{k=1}^{K}\log p(w_{i_k}|D\backslash\{D_I\};y_{i_k};\theta),$$

where D_I is a set of the masked tokens of the input sequence, w_{i_k} is a predicted token, K is a number of the masked tokens, y_{i_k} is the rhetorical relation corresponding to the masked token w_{i_k}, and θ is a model's parameters.

To perform this modified pre-training scheme, the SE layer is turned into the *discourse embedding* layer learned during the training procedure. The SE layer in the original BERT model is trained on the binary labels that denote whether the tokens belong to the first or the second sentence. In the proposed modified scheme, the segment layer should encode more than two labels that require its retraining to the size compatible with the number of discourse relations found in the text by discourse parsers. The corresponding layer is retrained from scratch. The architecture of the BERT model with discourse extension (disBERT) is presented in Figure 2.

The introduced disBERT operates with the original input sequence of tokens augmented with the relevant discourse relations retrieved from the DT. However, this information is too broad for the introduced discourse-aware MLM. In our research, the input data and the discourse structure of this input data need to be prepared in a format that the model could process. We propose an algorithm to represent the discourse structure of the text as the list of triples, where the relevant rhetorical relation connects the EDUs or the subset of several EDUs.

To retrieve these triples, we, first, build the dependency discourse graph (DDG) based on the DT, and then retrieve the desired triples directly from the graph.

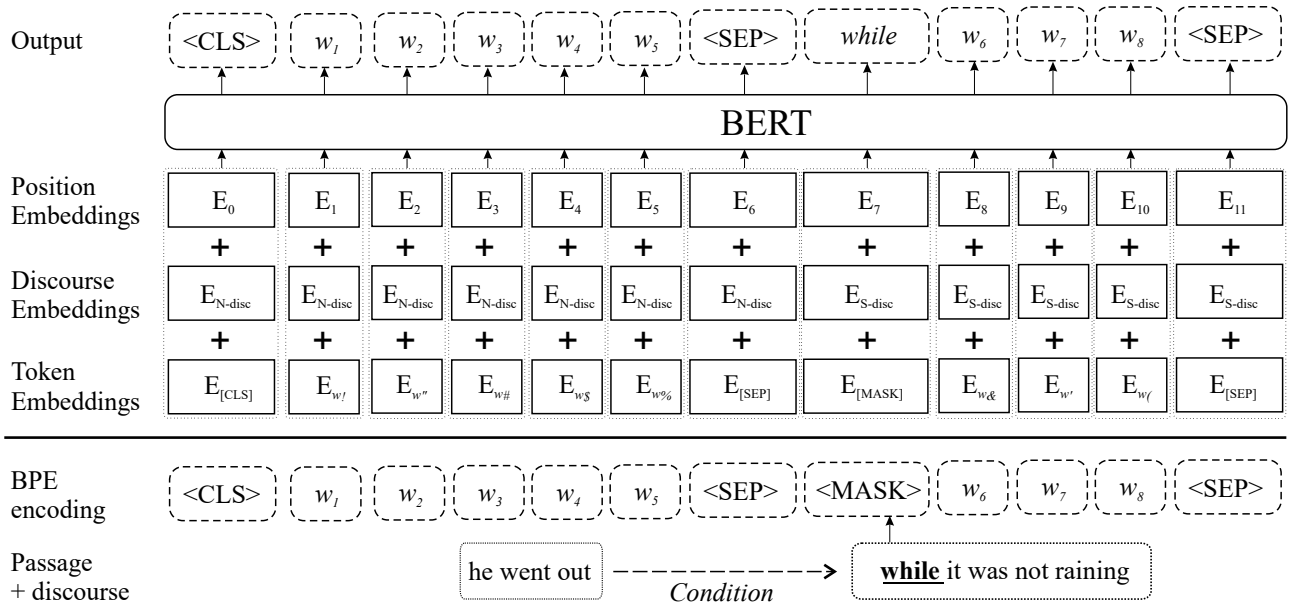

Figure 2 — Architecture of the modified discourse-enhanced disBERT model. The Figure introduces the modified BERT model extended with discourse structure. The *Position Embeddings*, *Token Embeddings*, and *Byte-Pair Encoding* are the standard layers of the original BERT. In contrast, the *Discourse Embeddings* layer is a modified layer presenting the encoding of the discourse relation that connects the nucleus EDU and satellite EDU divided by special [SEP] tokens in the input to the disBERT model.

The construction of a DDG begins with the empty graph, and then a DT is traversed bottom-up to retrieve the so-called head nodes for each node in the tree. These head nodes act as the vertices in the constructed graph, and every relevant rhetorical relation existing between the i^{th} and j^{th} EDU provides the direct edge in the graph, labeled with the name of the corresponding rhetorical relation. In more detail, a head node is chosen per the following rules:

- Head nodes for leaves are the leaves themselves.
- If a non-terminal node has nucleus and satellite children, also called an anti-symmetric relation, then the head is its nucleus child because the satellite units reflect the subordinate text spans and depend on the nuclei.
- If both children of a non-terminal node are the nuclei, also called a multi-nuclear relation, then the head node is the head of the left child

of the current node. The left child is chosen because the initial text is parsed left to right.

An example of DT → DDG transformation is shown in Figure 3.

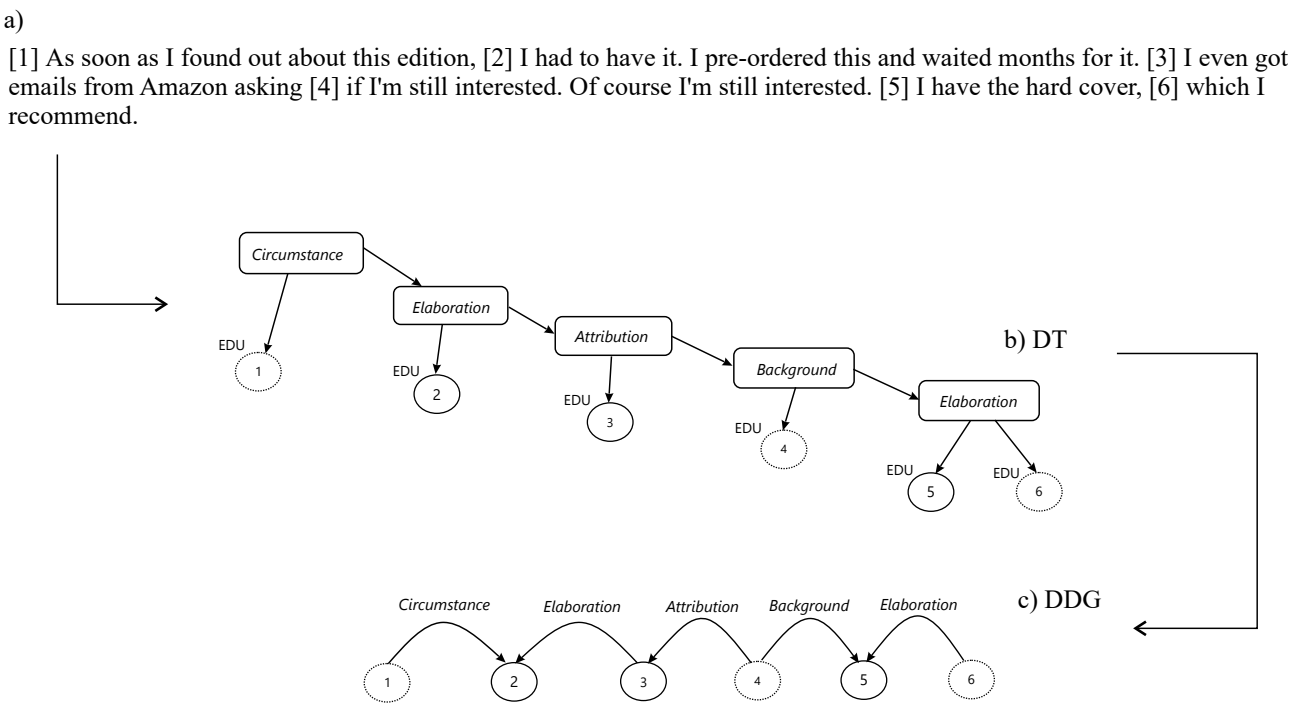

Figure 3 — The scheme for constructing the DDG from the DT. The text is split into 6 EDUs (a) and parsed into an RST DT (b). The obtained DT is then converted into the DDG (c). The solid lines denote nucleus nodes, while the dashed lines define satellites. Relations are given in italic.

Experimental evaluation. First, to check the model's awareness of discourse we challenge it with the *discourse connective prediction* task [29]. Showing that it outperforms the existing linguistic-free approaches by almost 7%.

Then, we launch disBERT model on the binary argument classification task, where in order to understand the argumentation structure of the text, the discourse structure should be considered.

For the experimental evaluation we utilize the dataset of Amazon users reviews (AR dataset [39]), and the UKP corpus [48]. As the baselines we used BiLSTM [48], MARGOT [31], MARGOT + TF-IDF [39], MARGOT + BoW, MARGOT + Discourse, BERT$_{base}$, BERT + Discourse [7], BERT-base$_{topic}$ presented in the studies for the same tasks.

We report F_1 score, precision, and recall as the metrics for all models. A shift-reduce discourse parser is used to extract relations and identify the nuclearity of the text spans [27][3] to construct the discourse tree for each of the documents.

The model performance on both AR and combined UKP corpus is shown in Table 1. Table 2 shows the results of the introduced disBERT model on each of the eight topics constituting the UKP dataset. This experiment has been done to assess the domain dependence of the presented model and to analyze whether the performance significantly drops for some topics, especially for the topic that was excluded from the training dataset.

Results on UKP Corpus. The BiLSTM baseline has achieved the worst classification performance with word2vec embeddings. The results show that even standard BERT_{base} significantly improves F_1 score, precision, and recall for AC in comparison to BiLSTM, while still, we should notice that recall that accounts for 0.26 is not high. Linguistic-aware models perform much better on the UKP corpus; incorporating discourse information directly into the model (disBERT) allows it to improve the recall by around 27pp compared to traditional BERT_{base}, which is comparable with topic-aware BERT. The pipeline of the BERT-based classifier and XGBoost trained on one-hot encoded rhetorical features slightly boosts traditional BERT_{base}; however, it does not outperform the linguistic-aware models.

Overall, the performance of the disBERT is slightly better than the one achieved by BERT_{topic} trained explicitly on the UKP corpus, by around 4pp and 2pp for precision and F_1 score, respectively.

Results on AR. The performance of the BERT-based models for the AR dataset is significantly higher than that for the UKP corpus. The AR dataset consists of more data. Therefore, we suggest that the BERT model was better trained on this data than on the smaller UKP Corpus. The traditional BERT_{base} model outperforms the stand-alone MARGOT; however, the MARGOT enlarged with the additional features performs better than all observed BERT-based models. We account for the fact that the Debater dataset that has been used for MARGOT training consists of human debates, including persuasive arguments on complex topics is close to the Amazon reviews constituting the AR dataset.

[3]The model for discourse parsing is available on https://github.com/jiyfeng/DPLP

Dataset	Model	Precision	Recall	F_1 score
UKP	BiLSTM [48]	0.41	0.16	0.23
	$BERT_{base}$	0.55	0.26	0.35
	BERT w. discourse [7]	**0.57**	0.32	0.41
	BERT-$base_{topic}$ [44]	0.53	0.52	0.52
	disBERT	0.56	**0.53**	**0.54**
"Movies and TV"	MARGOT [31]	0.54	0.77	0.63
	MARGOT (TF-IDF) [39]	0.73	0.78	0.75
	MARGOT w. BoW	0.74	**0.77**	0.75
	MARGOT w. disc.	0.76	0.78	0.78
	$BERT_{base}$	0.62	0.68	0.65
	BERT w. discourse [7]	0.65	0.69	0.67
	disBERT	**0.75**	0.73	**0.76**
"Electronics"	MARGOT [31]	0.53	0.74	0.61
	MARGOT (TF-IDF) [39]	0.65	0.68	0.66
	MARGOT w. BoW	0.74	0.77	0.75
	MARGOT w. disc.	**0.77**	0.71	0.74
	$BERT_{base}$	0.64	0.61	0.6
	BERT w. discourse [7]	0.62	0.63	0.62
	disBERT	0.71	**0.84**	**0.77**
"CDs and Vinyl"	MARGOT [31]	0.54	0.77	0.64
	MARGOT (TF-IDF) [39]	0.75	0.8	0.77
	MARGOT w. BoW	**0.74**	**0.8**	**0.77**
	MARGOT w. disc.	0.76	0.7	0.73
	$BERT_{base}$	0.62	0.68	0.65
	BERT w. discourse [7]	0.65	0.69	0.67
	disBERT	0.72	0.69	0.70
AR Dataset (combination of three categories)	disBERT	**0.83**	**0.80**	**0.79**

Table 1 — Experimental results assessing the disBERT model performance on the three categories of AR dataset and UKP corpus. The AR Dataset, a combination of three categories, corresponds to the experiment, when the model has been trained directly on the mixed texts from three types. Therefore, the results obtained for the UKP Corpus are more representative of AM domain than the results for the AR dataset.

Topic	Precision	Recall	F_1 score
Abortion	0.59	0.53	0.56
Cloning	0.62	0.57	0.59
Death penalty	0.68	0.66	0.67
Gun control	0.64	0.63	0.63
Marijuana legalization	0.62	0.63	0.62
Minimum wage	0.59	0.55	0.57
Nuclear energy	0.66	0.61	0.63
School uniforms	*0.67*	*0.55*	*0.60*

Table 2 — Experimental results on cross-topic evaluation. The model has been trained on 7 out of 8 topics and tested for each eight topics separately. The *'School uniforms'* topic has been excluded from the train set.

We compared the results on texts belonging to three categories from the AR dataset. Overall, the model's performance is almost the same for all three categories. We can notice that it outperforms all the other models by F_1 score for the "Movies and TV" and "Electronics" categories. For "CDs and Vinyl," the ensemble model combining MARGOT and BoW features slightly outperforms disBERT by 2pp in precision and 11pp in the recall. The fine-tuned disBERT shows improvement over the BERT_{base}. Precision boost accounts for 10pp, and recall boost is only 1pp. The disBERT boosts the performance of the ensemble model, which is on par with the results for the UKP corpus.

In contrast to the results on the topic-based UKP corpus, we noticed that discourse features are not as significant for the AR dataset. We explain it by the fact that the AR dataset was not initially collected for the AM tasks. It is more likely, but not necessary that the users label a review with argumentation components as applicable. Thus, assessing the usefulness of the reviews is the task connected to AC. However, it is not fully covered by AM.

The **fifth chapter** is dedicated to another model that we have proposed for incorporating discourse structure while solving more complex NLP tasks. The second approach refers to the discourse-aware attention mechanism inspired by the research presented in [59]. We utilize the constructed DDG and build the additional discourse-aware SAN that can narrow down the attention mechanism to the more connected discourse-aware components of the input paragraph. The experimental evaluation has proven that utilizing a discourse-aware attention mechanism is beneficial for the MRC launched on the lengthy and detailed-riddled passages.

The former scheme of the data augmentation procedure and retraining of the segmentation embedding layer with the discourse-conditioned MLM task has proven its applicability to argument classification. However, it also introduces some limitations to the model due to the specificity of the data pre-processing procedure.

The disBERT model operates on the discourse triplets. First, the initial text must be split into the smaller discourse-enhanced parts and be considered separately. However, it is not immediately evident how to implement this procedure for more complicated tasks, such as MRC, text summarization, or QA. These tasks also require discourse structure to be analyzed; however, the models need to process the lengthy passages in one step without splitting them into smaller sub-parts. A model needs to be able to process a long sequence of tokens to perform this task successfully.

Encapsulation of the discourse structure in MRC and summarization models may boost the model's performance by providing more complex mapping among the text spans via the attention mechanism. This chapter explores if and how discourse-level features, such as discourse relations connecting the text spans, fed to a neural MRC model on top of syntactic and semantic features or independently, can help answer complex, lengthy, multi-sentence questions [15]. We intend to develop a neural method that selects relevant words by only considering the related subset of words by analyzing syntactic, semantic, and discourse-level importance. A self-attention network (SAN) enriched with the discourse features – such as *Explanation* and *Condition* – retrieved from a text and combined with the classical transformer encoder are used to build linguistically-enhanced text representation to provide feature encoding. The overall model architecture is presented in Figure 4.

Experimental evaluation. This work relies upon four QA datasets with long, complex, multi-hop questions to observe if and how syntactic, semantic, and discourse-level features help provide the correct answers. The fine-tuned BERT model is used as the baseline. Additionally, the performance of the proposed system is compared with current state-of-the-art results published or obtained from the leaderboard for the corresponding dataset.

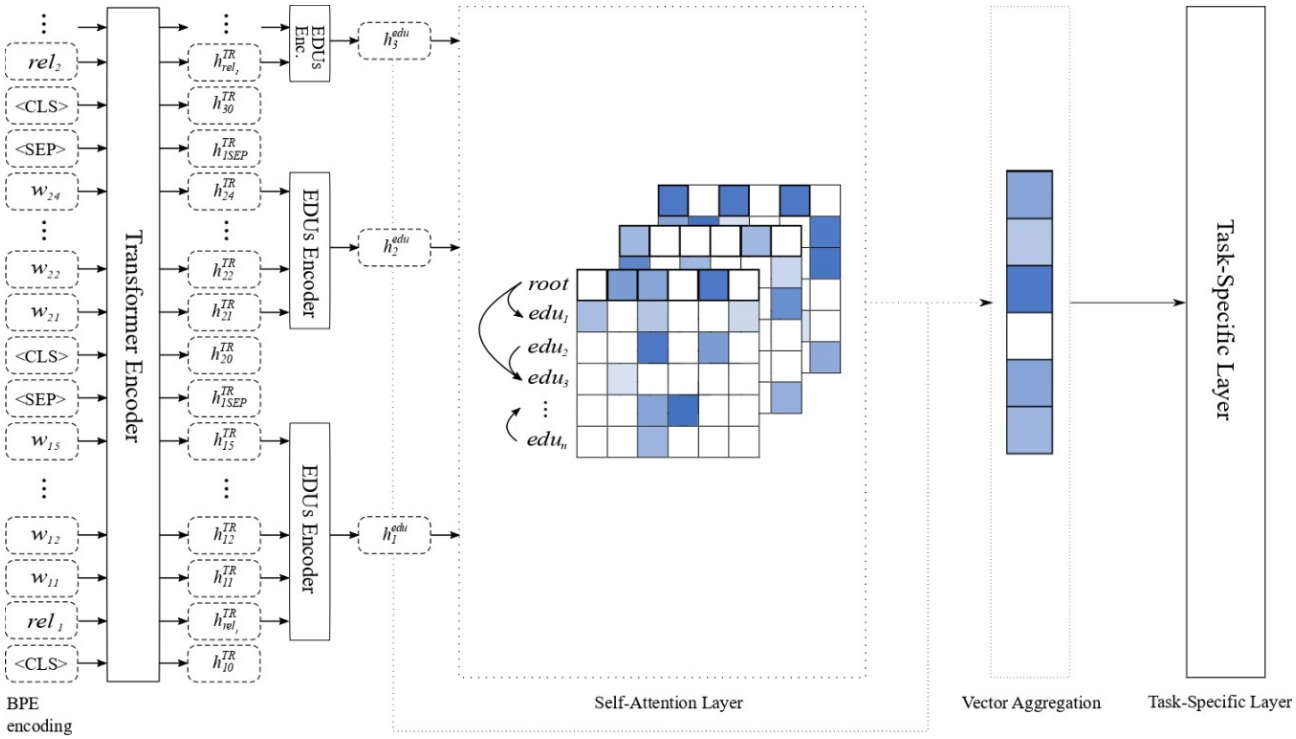

Figure 4 — Architecture of the MRC model extended with discourse-aware self-attention layer.

These experiments were divided into two parts to assess the influence of different linguistic features on the model performance. The results for the SQuAD datasets are provided in Table 3, and then the evaluation of the datasets with a more complex question design – NewsQA, QuAC, and MSQ – is presented in Table 4. The F_1 score is calculated as the harmonic average between precision and recall in all experiments. The results achieved by our MRC model are presented in the bottom block of the table. The results of the state-of-the-art models presented in the literature or public leaderboards for the available datasets are shown in the upper block; the * symbol denotes unpublished works. The results achieved by the MRC models relying on discourse information are in bold.

SQuAD. The performance on both SQuAD 1.1 and 2.0 test data is shown in Table 3. The default MRC baseline employs neither syntactic nor semantic information; this is a typical fine-tuned case BERT used as the encoder for the question and the passage. Moving towards syntactic, semantic, and discourse levels shows an average performance gain of 2.2, 3.4, and 3%, respectively. The improvement of the integrated system is 5.4%. We should also notice that the additional syntactic structure incorporation does not significantly boost the models performance that corresponds to the results of the BERT syntactic probing

Dataset/settings	v1.1 test	v2.0 test
	$\mathbf{F}_1$	$\mathbf{F}_1$
SQuAD leaderboard		
FPNet*	-	93.18
Retro-Reader [**Zhang2020_Retrospective**]	-	92.98
ALBERT [**Albert2020**]	-	92.20
LUKE*	95.4	-
Baseline	88.61	83.98
Syntax MRC	89.90	87.13
Semantic MRC	90.60	88.76
Discourse MRC	**90.08**	**88.60**
Syntax w. semantic w. discourse MRC	**93.14**	**90.20**

Table 3 — F_1 scores (%) on SQuAD 1.1 (v1.1) and SQuAD 2.0 (v2.0) datasets. The last row of the table represents the model combined with all three analyzed type of the linguistic information: syntax, semantics, and discourse.

Dataset/settings	NewsQA	QuAC	MSQ
	$\mathbf{F}_1$	$\mathbf{F}_1$	$\mathbf{F}_1$
literature + QuAC leaderboard			
SpanBERT [28]	73.6	-	-
DecaProp [49]	66.3	-	-
RoR*	-	74.9	-
FlowQA [23]	-	64.1	-
Baseline	66.48	65.69	60.66
Syntax MRC	70.95	71.09	66.79
Semantic MRC	71.84	70.15	66.55
Discourse MRC	**72.13**	**72.40**	**67.80**
Syntax w. semantic w. discourse MRC	**75.05**	**74.88**	**71.65**

Table 4 — F_1 scores (%) on complex questions datasets. The performance of other MRC models on MSQ dataset has not been published yet.

that has revealed the syntactic awareness of the BERT-based models. Although the introduced model could not outperform the best single and ensemble models, such as ALBERT and FPNet, it does boost the default linguistic-free baseline.

Complex datasets. Table 4 shows the result of the NewsQA, QuAC, and MSQ datasets. We combined these datasets together as the questions presented there are more complex than the one presented in SQuAD and require the reasoning over multiple sentences to retrieve the answer. Overall performance drops up to 20% when evaluating the datasets with more complex questions; analogously to Table 3, the default MRC employs no additional linguistic information. While the absolute performance value is lower than in Table 3, the performance boost due to linguistic information is greater. The average contributions of syntactic, semantic, and discourse levels are 5.3, 5.2, and 6.5%, respectively. The contribution of discourse-level features is the highest in this evaluation domain of MSQ. The improvement of the integrated system is almost 11% for MSQ and 9.5% on average. These results show that the longer and more complex the questions are, the higher the impact of linguistic information, especially at the discourse level. It should also be mentioned that the introduced ensemble model outperforms both the standalone fine-tuned BERT and current state-of-the-art models for NewsQA and achieves comparable results on QuAC.

In **chapter 6**, the explainability pipeline based on the discourse structure analysis is introduced. This research proposes an explanation pipeline that will provide us with the text spans – or *rationales* – retrieved from the input document with respect to its discourse structure that explain the model's decision. The rationales provided by the DL model serve to explain the model's decision. These text spans should be *faithful*, i.e. the model should rely on these text spans during decision making, and they should be *interpretable* by a human, i.e. the rationale needs to be presented in the format that a human can easily understand. The rationales obtained with the introduced pipeline are faithful, as they reflect the information utilized by a model to provide the decision. Additionally, they are interpretable by a human, as they are presented as text spans written in natural language and extended with the corresponding discourse structure that a human can understand.

Following prior work on the interpretability of language models [25], we propose an independent explanation pipeline used to obtain the rationales that

could potentially explain the model's decision and further check the influence of these rationales on the model's performance. The introduced pipeline consists of two independent components: extraction, which is responsible for extracting the inputs relevant to the model, and prediction, which tests the faithfulness of the obtained rationales using the initial model. The first part is supported by the interpretable ML method that operates on the discourse level of the text retrieved by the extraction block.

The proposed decomposition procedure reduces the training difficulty of previous explanation techniques, where extraction and prediction components have been trained jointly. This approach required optimization of the complex objective function using reinforcement algorithms. Besides, the obtained rationales are supported with the human-understandable discourse features that match the hidden text representation provided by the transformer-based model on the high-level linguistic attributes reflecting the linguistic structure encoded into the constructed embeddings.

The rationale extraction component (EC) builds the *rationale* candidates based on the model's inner structure (such as attention weights), which are further mapped to the relevant discourse structures identified by the formal context analysis (FCA) theory. The FCA-based method enables the model to retrieve the discourse structure of the text common for the texts belonging to different classes during the classification procedure. The pipeline in its entirety is presented in Figure 5.

The goal for rationales extraction is defined as follows: let the rationale be a specific combination of tokens retrieved from the input sequence, $rat = \{w_1, w_2, ..., w_K\}$, where $w_k \in W$ and $k \in I$; W is an input sequence of tokens, and I is a set of indices referring to the informative text spans. Ideally, the combination of input tokens that provides the minimum value of the loss function for the analyzed task should be found; however, it requires training the model with different input token combinations. Unfortunately, identifying the subsets of tokens that lead to minimal loss is resource-intensive and time-consuming. Additionally, the obtained rationales should explain the model's decision in a way that will be understandable by a human. We predict that when the grammatical consistency of the found rationales is high, there will be substantial explainability improvement.

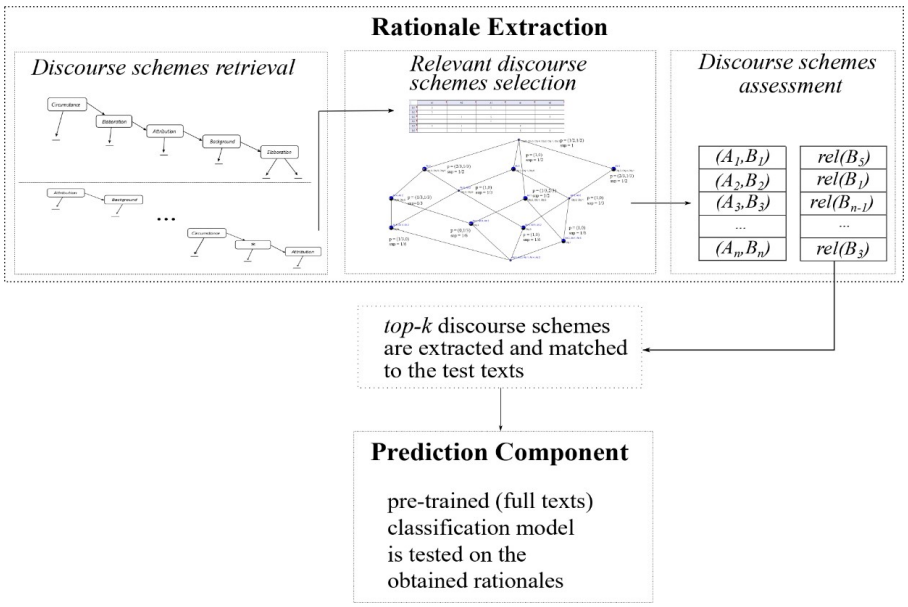

Figure 5 — *Independent Explanation Pipeline (IEP)* for rationales construction. Rationale Extraction (RE) component is used to provide the rationales candidates based on the relevant discourse features; Prediction Component (RP) is used for constructed rationales validation.

The EC of the IEP serves to retrieve relevant discourse features that might be used for explaining the model's decision. The text itself is not of interest during relevant discourse features extraction, only its discourse structure. It is retrieved as the sub-graphs of the initial DT, or DDG, where the specific text span constituting some terminal node, referring to the EDU, is replaced by the asterisk symbol. In contrast, non-terminal nodes are labeled with the name of discourse relation.

Thus, the task of rationale, or rat, extraction turns to the task of assessing the $rel(w_i|DDG_i)$ for each $i \in N$, where $rel(w_i|DDG_i)$ is a relevance score of the token w_i conditioned on its discourse role in the text, and DDG_i is a frequent sub-graph of the initial DDG retrieved by the *gSpan* algorithm [55]. It is assumed that for discourse-dependent tasks without significant loss, the $rel(DDG_i)$ can be evaluated, and then the w_i corresponding to the $rel(DDG_i)$ can be chosen. We support this assumption with the experimental evaluation performed on several NLP tasks, where the discourse structure has been proven to be important.

To assess the relevancy of the discourse sub-graphs, we utilize the FCA theory. The FCA-based model introduced in [18] is utilized to select the informative features, providing an interpretable method for analyzing the groups of objects and the features they share. In more detail, the first step of this introduced pro-

cedure is to one-hot encode the DDG representation of the input text to get its feature description. FCA theory is then used to build the concept lattice and find the hypothesis that reflects which feature combination is relevant for identifying the argumentativeness of the text.

The FCA-based classification reflects the similarity of the concepts rather than their statistical representation. Thus, the similarity of graph structure can be analyzed without considering the exact words utilized in the classification procedure. We have introduced the classification pipeline with the FCA-based technique in reference [19] operating with the binary features, and trying to identify the most relevant feature subset based on their influence on the classification performance. In this work, we leverage the presented approach for the other NLP domain, still operating with the objects described by the binary features.

The representation of each document $x \in X$ is retrieved from its DDG. The common sub-graphs are extracted from the DDGs and encoded as the binary features so that a text is represented as a binary vector in this feature space. The text is assigned to a binary feature if the subgraph is found in the initial DDG describing the observed document. Finally, we apply the approach introduced in [19] to retrieve the subset of relevant binary features – or frequent discourse sub-graphs – and map them to the obtained rationale candidates retrieved by EC.

The whole IEP can be described as follows:

1. Obtain the rationales as the set of the text spans obtained with respect to the informative discourse components.
2. Provide the model's prediction based on the obtained rationales.
3. Compare the obtained results with the ones of simple attention and regression component obtained from the standard procedure of assessing the rationales using the pure transformer-based models.

Experimental evaluation. The experimental evaluation was performed on different classification NLP datasets: two datasets for AC (AR and UKP Corpus), two sentiment classification datasets (Stanford Sentiment Treebank (SST) [47], Movies [57]), and AgNews [10] for multi-class classification.

We use the SST dataset, where the documents are split into two classes: positive and negative. There are 9,613 documents available for fine-tuning. Movies is a dataset of users' reviews labeled by the corresponding sentiment mark. This

dataset consists of 1,999 documents. AgNews dataset is a multi-class classification dataset, where the news articles are labeled with one of the corresponding topic: *Sport*, *Science*, *Business*, and *World*. This dataset contains 127,600 documents.

Standard BERT$_{base}$ and *disBERT* and BERT extended with discourse attention are used as the baseline for retrieving the rationales. We predict that the discourse-aware models should be more attentive to the constructed rationales, as they were trained to capture the discourse structure during fine-tuning procedures.

We report the results for rationales construction obtained with the reinforcement algorithm proposed by Lei et al. [30], Bastings et al. model[4] [2] that utilizes reparametrization trick for the rationales extraction, and an attention-based FRESH model [25] (**Att.-based**) that outputs *top-k* tokens with the highest attention scores as the rationales. We also report the performance of the models on the full texts (**Full text**) in order to assess the faithfulness of the constructed rationales. The baselines' performance is reported from the corresponding papers.

Table 5 presents the results obtained by different models on the rationales constructed using various methods for the analyzed datasets.

PC (model)	RE (approach)	AR	UKP	SST	Movies	AGNews
BERT$_{base}$	Full text	0.60	0.35	0.90	0.94	0.96
	Lei et al.	0.52	0.33	0.74	**0.92**	0.87
	Bastings et al.	0.51	0.28	0.59	0.72	–
	Att.-based	0.63	0.32	**0.81**	0.91	**0.94**
	IEP	**0.64**	**0.34**	0.71	0.80	0.82
disBERT	Full text	0.68	0.54	0.67	0.74	0.76
	Lei et al.	0.52	0.45	0.54	0.62	0.67
	Bastings et al.	0.51	0.48	0.60	**0.65**	—-
	Att.-based	0.63	0.52	**0.61**	0.59	**0.68**
	IEP	**0.65**	**0.53**	0.53	0.57	0.62
BERT ext. with discourse-aware SAN	Full text	0.77	0.69	0.89	0.87	0.85
	Lei et al.	0.62	**0.65**	0.74	**0.92**	**0.87**
	Bastings et al.	0.69	0.57	0.59	0.52	-—
	Attn-based	0.53	0.52	0.71	0.71	0.82
	IEP	**0.72**	0.63	**0.75**	0.67	0.81

Table 5 — Prediction component performance on different datasets. F_1 scores of the correct prediction are presented in the table.

A comparison is provided via two perspectives: the rationale extraction approaches and the explainability of different models. Thus, whether the IEP influences only the discourse-aware models or also explains the standard vanilla-based transformers is analyzed.

[4]https://github.com/bastings/interpretable$_p$*redictions*

The performance of the *disBERT* and the discourse-aware attention model using the rationales is almost comparable to the performance of these models when tested on the full texts. The models' performance on rationales drops on sentiment analysis datasets by up to 17pp for the Movies reviews, suggesting that IEP is more suitable for discourse-aware models designed for complex tasks. We can also notice that the introduced IEP in the combination with disBERT model achieves the highest F_1 score in comparison to the other extraction methods. The performance of the models tested on the shortened texts consisting of only the constructed rationales is not significantly lower than the one achieved for the full texts (e.g., 0.68 vs 0.65 for disBERT model combined with the IEP). This indicates that the constructed rationales are faithful and encode almost all the relevant information from the initial text. The combination of the discourse-aware IEP and the standard BERT_{base} model outperforms BERT_{base} on the AR dataset shows that the RE component retrieved only the informative text spans and removed the irrelevant parts of the text.

Finally, **conclusion** summarizes the dissertation, outlining the main findings of the conducted research.

CONCLUSION

The main results of the presented dissertation are as follows:

1. The discourse awareness of the pre-trained models is analyzed.
2. Two modifications for the existing pre-trained BERT model are proposed (disBERT and BERT w. discourse-aware attention mechanism).
3. The experimental evaluation of the proposed models is presented showing their applicability to the discourse-aware NLP tasks, such as AC and MRC.
4. Both the introduced models outperform the original BERT model on AC and MRC tasks.
5. disBERT model is checked on discourse probing task outperforming the existing transformer-based models.
6. The models are fine-tuned and made publicly available for the community.
7. The independent explainability pipeline combining the FCA-based method with the pre-trained transformer-based models is presented for generating the rationales for the analyzed tasks.

References

[1] Yossi Adi et al. "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 2017.

[2] Joost Bastings, Wilker Aziz, and Ivan Titov. "Interpretable neural predictions with differentiable binary variables". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020. DOI: 10.18653/v1/p19-1284.

[3] Yonatan Belinkov et al. "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 1–10. URL: https://aclanthology.org/I17-1001.

[4] Yonatan Belinkov et al. "What do neural machine translation models learn about morphology?" In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. 2017. DOI: 10.18653/v1/P17-1080.

[5] Terra Blevins, Omer Levy, and Luke Zettlemoyer. "Deep RNNs encode soft hierarchical syntax". In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 2. 2018. DOI: 10.18653/v1/p18-2003.

[6] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. "Inducing relational knowledge from BERT". In: *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. 2020. DOI: 10.1609/aaai.v34i05.6242.

[7] Tuhin Chakrabarty et al. "AmperSand: Argument mining for persuasive online discussions". In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/d19-1291.

[8] Junyoung Chung et al. "Gated feedback recurrent neural networks". In: *32nd International Conference on Machine Learning, ICML 2015*. Vol. 3. 2015.

[9] Kevin Clark et al. “What Does BERT Look at? An Analysis of BERT’s Attention”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828.

[10] Gianna M. Del Corso, Antonio Gullí, and Francesco Romani. “Ranking a stream of news”. In: *14th International World Wide Web Conference, WWW2005*. 2005. DOI: 10.1145/1060745.1060764.

[11] Daniel Cristea, Nancy Ide, and Laurent Romary. “Veins Theory : A Model of Global Discourse Cohesion and Coherence”. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 18th International Conference on Computational Linguistics COLING98ACL98* 1 (1998).

[12] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2019.

[13] Boris Galitsky. “Learning Discourse-Level Structures for Question Answering”. In: *Developing Enterprise Chatbots* (2019), pp. 177–219. DOI: 10.1007/978-3-030-04299-8_7.

[14] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. “On a Chatbot Conducting Dialogue-in-Dialogue”. In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, Sept. 2019, pp. 118–121. DOI: 10.18653/v1/W19-5916. URL: https://aclanthology.org/W19-5916.

[15] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. “Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sept. 2021, pp. 444–453. URL: https://aclanthology.org/2021.ranlp-1.51.

[16] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. "Relying on Discourse Trees to Extract Medical Ontologies from Text". In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 215–231. ISBN: 978-3-030-86854-3. DOI: 10.1007/978-3-030-86855-0_15. URL: https://doi.org/10.1007/978-3-030-86855-0_15.

[17] Boris A. Galitsky, Sergei O. Kuznetsov, and Daniel Usikov. "Parse thicket representation for multi-sentence search". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7735 LNCS. 2013. DOI: 10.1007/978-3-642-35786-2_12.

[18] Bernhard Ganter and Rudolf Wille. "Formal concept analysis". In: Springer Verlag, Berlin, 1998.

[19] Elizaveta Goncharova and Sergei Kuznetsov. "Increasing the Efficiency of Packet Classifiers with Closed Descriptions". In: *In proceedings of the 7th International Workshop "What can FCA do for Artificial Intelligence"? co-located with International Joint Conference on Artificial Intelligence, FCA4AI@IJCAI*. 2019, pp. 75–86.

[20] Quentin Grail, Julien Perez, and Eric Gaussier. "Globalizing BERT-based transformer architectures for long document summarization". In: *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. 2021. DOI: 10.18653/v1/2021.eacl-main.154.

[21] André Grüning and Andrej A. Kibrik. "Modelling referential choice in discourse: a cognitive calculative approach and a neural network approach". In: *Anaphora processing: linguistic, cognitive and computational modelling* (2005).

[22] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. "Fast Axiomatic Attribution for Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 19513–19524.

[23] Hsin Yuan Huang, Wen Tau Yih, and Eunsol Choi. "FlowQA: Grasping flow in history for conversational machine comprehension". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.

[24] Beltagy Iz, Peters Matthew E., and Cohan Arman. "Longformer: The Long-Document Transformer". In: *arXiv* (2020).

[25] Sarthak Jain et al. "Learning to Faithfully Rationalize by Construction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4459–4473. DOI: 10.18653/v1/2020.acl-main.409. URL: https://aclanthology.org/2020.acl-main.409.

[26] Ekaterina Jasinskaja, Jörg Mayer, and David Schlangen. "Discourse Structure and Information Structure: Interfaces and Prosodic Realization". In: *Interdisciplinary Studies on Information Structure (ISIS)* (2004).

[27] Yangfeng Ji and Jacob Eisenstein. "Representation learning for text-level discourse parsing". In: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. Vol. 1. 2014. DOI: 10.3115/v1/p14-1002.

[28] Mandar Joshi et al. "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *Transactions of the Association for Computational Linguistics* 8 (2020). DOI: 10.1162/tacl_a_00300.

[29] Fajri Koto, Jey Han Lau, and Timothy Baldwin. "Discourse Probing of Pretrained Language Models". In: 2021. DOI: 10.18653/v1/2021.naacl-main.301.

[30] Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. DOI: 10.18653/v1/D16-1011. URL: https://aclanthology.org/D16-1011.

[31] Marco Lippi and Paolo Torroni. "MARGOT: A web server for argumentation mining". In: *Expert Systems with Applications* 65 (2016). DOI: 10.1016/j.eswa.2016.08.050.

[32] Nelson F. Liu et al. "Linguistic knowledge and transferability of contextual representations". In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Hu-*

man Language Technologies - Proceedings of the Conference. Vol. 1. 2019. DOI: 10.18653/v1/n19-1112.

[33] Annie Louis, Aravind Joshi, and Ani Nenkova. "Discourse Indicators for Content Selection in Summaization". In: Sept. 2010, pp. 147–156.

[34] T.P. Makhalova et al. "FCA-based approach for interactive query refinement with IR-chatbots". In: *In proceedings of the Russian Advances in Artificial Intelligence, RAAI 2020*. Vol. 2648. 2020, pp. 144–156.

[35] William Mann and Sandra Thompson. "Rhetorical Structure Theory: Toward a functional theory of text organization". In: *Text* 8 (3 1988). DOI: 10.1515/text.1.1988.8.3.243.

[36] Tobias Mayer, Elena Cabrio, and Serena Villata. "Transformer-based argument mining for healthcare applications". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence*. Vol. 325. 2020. DOI: 10.3233/FAIA200334.

[37] Tomaš Mikolov et al. "Recurrent neural network based language model". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2010. DOI: 10.21437/interspeech.2010-343.

[38] Pramod K. Mudrakarta et al. "Did the model understand the question?" In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. 2018. DOI: 10.18653/v1/p18-1176.

[39] Marco Passon et al. "Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining". In: *Proceedings of the 5th Workshop on Argument Mining*. 2018, pp. 35–39. DOI: 10.18653/v1/w18-5205.

[40] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global vectors for word representation". In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2014. DOI: 10.3115/v1/d14-1162.

[41] Matthew E. Peters et al. "Deep contextualized word representations". In: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Vol. 1. 2018. DOI: 10.18653/v1/n18-1202.

[42] Rashmi Prasad et al. "The Penn Discourse TreeBank 2.0." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

[43] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". In: *Preprint* (2018).

[44] Nils Reimers et al. "Classification and clustering of arguments with contextualized word embeddings". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/p19-1054.

[45] Xing Shi, Inkit Padhi, and Kevin Knight. "Does string-based neural MT learn source syntax?" In: *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2016. DOI: 10.18653/v1/d16-1159.

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *Workshop at International Conference on Learning Representations*. 2014.

[47] Richard Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2013.

[48] Christian Stab et al. "Cross-topic argument mining from heterogeneous sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. 2018. DOI: 10.18653/v1/d18-1402.

[49] Yi Tay et al. "Densely connected attention propagation for reading comprehension". In: *Advances in Neural Information Processing Systems*. Vol. 2018-December. 2018.

[50] Ian Tenney, Dipanjan Das, and Ellie Pavlick. “BERT rediscovers the classical NLP pipeline”. In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020. DOI: 10.18653/v1/p19-1452.

[51] Ian Tenney et al. “What do you learn from context? Probing for sentence structure in contextualized word representations”. In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.

[52] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*. 2018. DOI: 10.18653/v1/w18-5446.

[53] Alex Wang et al. “SuperGLUE: A stickier benchmark for general-purpose language understanding systems”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[54] Florian Wolf and Edward Gibson. “Representing Discourse Coherence: A Corpus-Based Study”. In: *Computational Linguistics* 32 (June 2005), pp. 249–287.

[55] Xifeng Yan and Jiawei Han. “gSpan: Graph-based substructure pattern mining”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2002. DOI: 10.1109/icdm.2002.1184038.

[56] Manzil Zaheer et al. “Big bird: Transformers for longer sequences”. In: *Advances in Neural Information Processing Systems*. Vol. 2020-December. 2020.

[57] Omar F. Zaidan, Jason Eisner, and Christine D. Piatko. “Using "annotator rationales" to improve machine learning for text categorization”. In: *NAACL HLT 2007 - Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*. 2007.

[58] Zhuosheng Zhang et al. “Semantics-Aware BERT for Language Understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020. DOI: 10.1609/aaai.v34i05.6510.

[59] Zhuosheng Zhang et al. “SG-Net: Syntax-Guided Machine Reading Comprehension”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05 2020). DOI: 10.1609/aaai.v34i05.6511.

[60] Wei Zhao et al. “Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance”. In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/d19-1053.