Pre-Trained Models. BERT

BERT is a transformer-based model pre-trained on two large corpora: BooksCorpus (800M words) and English Wikipedia (2,500M words).

Two pre-trained objectives:

- Masked Language Modeling (MLM)
  $L_{MLM}(D_I|D\backslash\{D_I\}) = \frac{1}{K}\sum_{k=1}^{K} \log p(w_{i_k}|D\backslash\{D_I\};\partial)$.
  D_I is a subset of masked tokens, K is a number of masked tokens.
- Next Sentence Prediction
  $P(NSP) = \texttt{softmax}(FC_2(tanh(FC_1(BERT^N_{[CLS]}))))$.

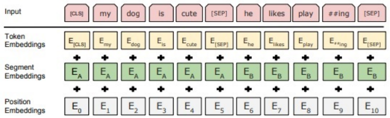


Figure: Input format for the BERT model

Linguistic Awareness of the Pre-Trained Models

Does BERT understand language?
(Niven et al., 2019)[2]
Conclusion: the model is trained to utilize some statistical patterns in the dataset and simple adversarial examples degrade its performance.

- Semantics [1, 3]
- Syntax [2]
- Discourse [4, 5]

[2] T. Niven and H.Y. Kao. *Probing Neural Network Comprehension of Natural Language Arguments. ACL, 2019* evaluate BERT on Argument Reasoning Comprehension.

Discourse Structure

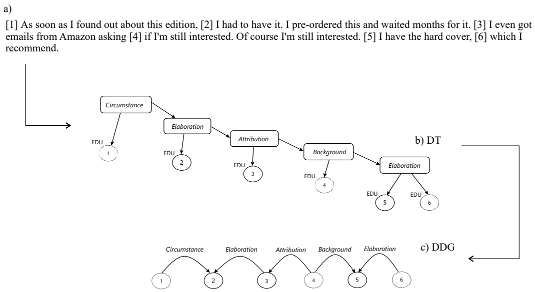
Rhetorical Structure Theory (RST) (Mann and Thompson, 1988)

Elementary Discourse Units (EDUs)

- Atomic text spans that are connected by the rhetorical relations.
- Nucleus (N) is a main span, satellite (S) is a peripheral span.

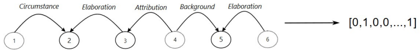
Rhetorical relations

- EDUs can be joint by the rhetorical relations which show the type of their connection to each other.
- The EDUs connected by the rhetorical relations are organized into the hierarchy constituting the discourse tree (DT).
- Construct dependency discourse graph (DDG).



Standard Approach to Discourse Encoding

- One-hot encoding of the discourse structure
  Does not consider parts of the text among which the relation exists.
- Graph neural network (GNN)
  Requires training of the additional neural network component.



DISCOURSE-AWARE BERT MODEL

Masked Language Modeling (MLM)

MLM pre-training

It was raining ***but*** *he went out.* [Contrast]
It was raining ***and*** *he went out.* [Elaboration]

MLM pre-training

It was raining [MASK] he went out $\rightarrow$ BERT $\rightarrow$ when

The prediction should depend on the rhetorical relation.

Example Inputs

Select a Sentence

Sentence

It was raining [MASK] he went out.

Run Model

Prediction	Score
It was rain ##ing **when** he went out .	64 %
It was rain ##ing **as** he went out .	22,1 %
It was rain ##ing **and** he went out .	5,7 %
It was rain ##ing **so** he went out .	1,6 %
It was rain ##ing **before** he went out .	1,5 %

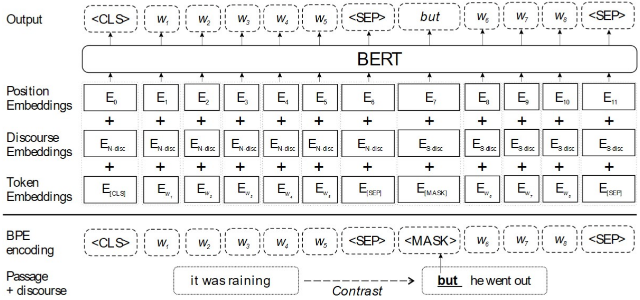
MLM with Discourse Extension

Discourse-aware MLM:

$p\left(\cdot|C_{\setminus w_t}, rel_t\right)$
C is a context,
w_t is a masked token,
rel_t is a rhetorical relation.

Triples extraction

The required representation of a document $d_k \in D$ extended with its discourse structure:



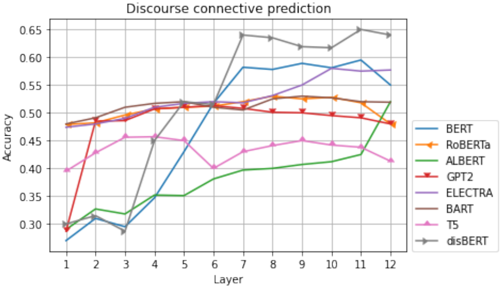
$$d_k = \{(edu_{11}^k, edu_{12}^k, rel_1^k), (edu_{21}^k, edu_{22}^k, rel_2^k), ..., (edu_{N1}^k, edu_{N2}^k, rel_N^k)\}$$

Discourse Probing

Discourse Connective Prediction

Sampled DisSent dataset (Nie et al., 2019)

15 labels
Split: 10K/1K/1K



Experimental Evaluation

Argument Classification task

Oracle labels:
0 – non-argumentative texts
1 – argumentative texts

Datasets:

UKP corpus is a topic-dependent AC dataset. There are 8 topics, train set includes 7 topics, test set covers all 8 topics.
Amazon Reviews (AR) consists of the users' reviews divided into 2 classes based on the usefulness of the reviews.

F_1-score is used for the assessment.

Experimental Results

Dataset	Model	Precision	Recall	F_1-score
UKP	BiLSTM	0.41	0.16	0.23
	BERT_{base}	0.55	0.26	0.35
	BERT w. discourse	**0.57**	0.32	0.41
	BERT-base_{topic}	0.53	0.52	0.52
	disBERT	0.56	**0.53**	**0.54**
"Movies and TV"	MARGOT	0.54	0.77	0.63
	MARGOT (tf-idf)	0.73	0.78	0.75
	MARGOT w. BoW	0.74	0.75	0.74
	MARGOT w. disc.	**0.75**	**0.78**	**0.76**
	BERT_{base}	0.62	0.68	0.65
	BERT w. discourse	0.65	0.69	0.67
	disBERT	**0.75**	0.73	0.74
AR Dataset (combination of three categories)	*disBERT*	**0.83**	**0.80**	**0.81**

Discourse Features Importance

Subsets of discourse features:

1. Elaboration, Circumstance, Background, Interpretation, Evaluation.
2. Attribution, Evidence, Example, Explanation, Reason, Consequence.
3. Cause, Result, Purpose, Means.
4. Contrast, Antithesis, Concession.
5. Condition, Comparison.
6. Same unit, Textual organization.
7. List, Sequence, Disjunction.

Dataset	Feature combination	Precision	Recall	F_1-score
AR Dataset (combination of three categories)	[1]	0.69	0.45	0.54
	[2]	0.74	0.53	0.62
	[3]	0.75	0.67	0.71
	[4]	0.72	0.6	0.65
	[5]	0.69	0.52	0.59
	[6]	0.73	0.71	0.72
	[7]	0.62	0.43	0.51
	[2,3,4,5]	**0.84**	**0.82**	**0.83**
	[1,6,7]	0.61	0.35	0.44

Application of the AC Model

A chatbot for navigating a user through the concept-based knowledge model built for the database of the online store items extended with their textual descriptions.

$G, (D, \sqcap), \delta$ is a pattern structure, where G is a set of objects, and $(D, \sqcap)$ is a semilattice of object descriptions; $\delta : G \to D$

$$A^{\square} = \sqcap_{g \in A}(\delta(g)), A \subseteq G$$
$$d^{\square} = \sqcap_{g \in A}(\delta(g)), d \in (D, \sqcap)$$

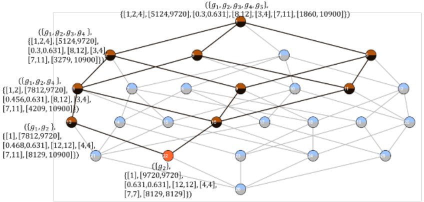
A pair (A, d) with $A^{\square} = d$ and $d^{\square} = A$ is a pattern concept.

Ordered pattern concepts $(A_1, d_1) \leq (A_2, d_2) \Leftrightarrow A_1 \subseteq A_2$ is a pattern concept lattice.

Application of the AC model

A pattern concept lattice built on the structural attributes of the item and the textual features derived from users' reviews is constructed to navigate a user to the desired set of items[3].

	Brand	Battery	Weight	Cam. Resolution	RAM	Screen Size	Price	Text description
g_1	1	7812	0.468	12	4	11	10900	good camera
g_2	1	9720	0.631	12	4	7	8129	fantastic photo
g_3	4	5124	0.3	8	3	10.5	3279	low battery cap.
g_4	2	8134	0.456	8	3	10.5	4209	too heavy
g_5	1	7000	0.44	8	3	10.5	1860	lack of autonomy



Data:

- 500 items from the *electronics* dataset, each item is assigned to 5-10 users' reviews (8960 reviews).
- Using *disBERT* model we reduced the number of the analyzed reviews up to 2478 argumentative ones.
- The total usefulness score has grown from 2.6 to 3.7.

[3] Goncharova et al. On a Chatbot Navigating a User through a Concept-Based Knowledge Model, EcomNLP 2020

Part 1. Summary

1. A discourse-aware *disBERT* model is presented. It allows one to encode discourse features into the standard BERT model via modified MLM pre-training task.
2. Experimental evaluation on AC task shows that *disBERT* outperforms standard BERT model.
3. The influence of various discourse features is analyzed.
4. DDG construction allows one to retrieve only relevant discourse features from the initial DT.
5. Splitting the input sequence into the triplets makes the model more suitable for classification tasks.

DISCOURSE-AWARE ATTENTION MECHANISM

Machine Reading Comprehension

The task is to retrieve relevant text spans from the passage that contain the answer to the passage. It is necessary to identify long-range dependencies existing among the text spans.
$< P, Q, A >$, where P defines a passage, Q is a question, and A is a correct answer.

Transformer-based model can be used to calculate vector representation (embedding) of the input sequence.

Input format: [CLS] P [SEP] Q [SEP]

The goal is to assess whether the tokens constituting a passage P should be included into the final answer A.

Case Study. Importance of the Oriented Attention

P: Viruses, bacteria, and fungi can all cause pneumonia. Common causes of viral pneumonia are influenza and respiratory syncytial virus. A common cause of bacterial pneumonia is Streptococcus pneumonia. However, clinicians are not always able to find out which germ caused someone to get sick with pneumonia.
Q: Who experience difficulties finding causes for pneumonia?

BERT_{base} answer is viruses, bacteria, and fungi.
The correct answer is clinicians.

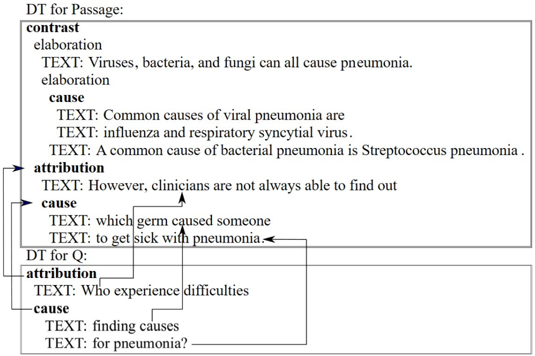
Wrong association:
Viruses, bacteria, and fungi $\leftrightarrow$ Who

Discourse-Aware MRC

A mapping between Q and P:

- Q: attribution → P: attribution
- Q: cause → P: cause
- Q: "causes" → P: "caused"

The model attends each word to the relevant text spans in the input passage.



Discourse-Aware Attention Model

Augmenting a text with discourse relations

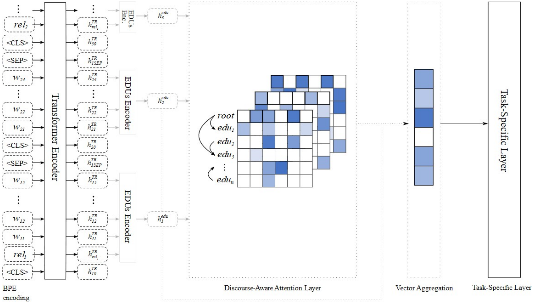
$S = \{s_1, s_2, \ldots, s_n\} \xrightarrow{\texttt{discourse parser}} \overline{S}_{rel} = \{rel_1, edu_{11}, edu_{12}, rel_2, \ldots, rel_m, edu_{m1}, edu_{m2}\}$

Discourse dependency of interest mask

P_i is a set of ancestor EDUs calculated for each edu_i and rel_i.

$$M[i,j] = \begin{cases} 1 & \text{if } j \in P_i \text{ or } j = i \\ 0 & \text{otherwise.} \end{cases}$$

Used in calculation of the attention weights [4].



[4] B. Galitsky, D. Ilvovsky, E. Goncharova. Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension. RANLP, 2021

Datasets and Setup

- SQuAD 2.0 dataset.
  100,000 QA pairs.
- NewsQA dataset.
  119,633 QA pairs
- QA in Context (QuAC) dataset.
  100,000 QA pairs.
- MSQ dataset.
  162,745 QA pairs.

$F_1 = 2\frac{recall \cdot precision}{recall + precision}$.

$precision = \frac{\#tokens_{common}}{\#tokens_{predicted}}$; $recall = \frac{\#tokens_{common}}{\#tokens_{gt}}$.

Baseline is fine-tuned BERT$_{base}$.

Experimental Results

Dataset/settings	v1.1 test F1	v2.0 test F1
SQuAD leaderboard		
FPNet*	-	93.18
Retro-Reader	-	92.98
ALBERT	-	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**

Dataset/settings	NewsQA F1	QuAC F1	MSQ F1
literature + QuAC leaderboard			
SpanBERT	73.6	-	-
DecaProp	66.3	-	-
RoR*	-	74.9	-
FlowQA	-	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**

Part 2. Summary

1. A discourse-aware attention mechanism is proposed.
2. Experimental evaluation on MRC task shows that the modified model outperforms the existing techniques on the MRC datasets with lengthy passages.

INDEPENDENT EXPLANATION PIPELINE

Explainability of the Pre-Trained Models

Rationale is a text span that can explain a model's decision.

Requirements:

- They should be utilized by the model to provide a decision.
- They should be easily understandable by a human.

Rationales Extraction

Possible solutions:

Construct an **Extractor** that retrieves the *rationales* and then train a **Classifier** to find out the correctness of the *rationales*.

Needs a lot of human-constructed rationales.

Independent Explanation Pipeline (IEP)

- Fine-tune the pre-trained model on the downstream task.
- Use some *scoring* technique to obtain the *rationales* candidates (attention weights).
- Choose *top-k* tokens as the *rationales*[5].

Drawbacks: The obtained rationales are usually not grammatically consistent.

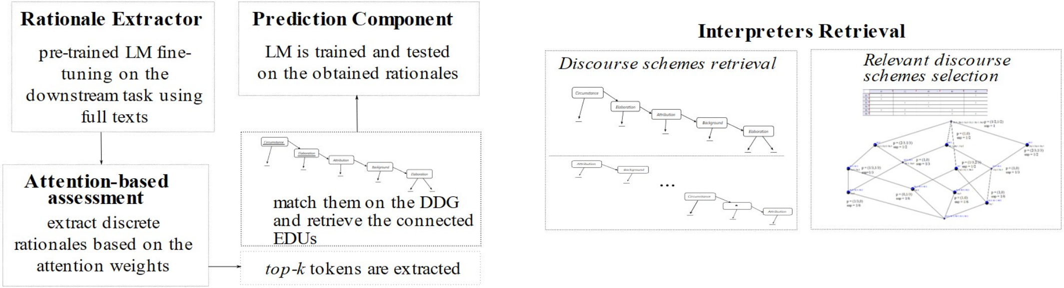
[5] S. Jain, S. Wiegreffe, Y. Pinter, and B.C. Wallace, Learning to Faithfully Rationalize by Construction, ACL, 2021

Independent Explanation Pipeline with Discourse Extension

Extract rationales as a part of the text that explains a model's decision.
$r = \{w_1, w_2, ..., w_l\}$, where $w_i \in D$
$r = \{edu_1, edu_2, ..., edu_l\}$, where
$edu_i = \{w_1^i, w_2^i, ..., w_{i_l}^i\}$, $w_j^i \in D$.

The rationales are extracted based on the assigned discourse schemes.
Pure discourse schemes are used to assess the relevancy of the texts for the classification

Independent Explanation Pipeline (IEP)[6].



6E. Goncharova, S. Kuznetsov, Increasing the efficiency of packet classifiers with closed descriptions. FCA4AI, 2019

Experimental Evaluation

Datasets:

Argumentation Classification:

- Amazon Reviews
- UKP Corpus

Sentiment Classification:

- SST-2 Binary Sentiment Classification Benchmark
- IMDB Movies Reviews
- AG News Sentiment Classification

Experimental Results

PC (model)	RE (approach)	AR	UKP	SST	Movies	AGNews
BERT_{base}	Full text	0.6	0.35	0.9	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	—
	Attn.-based (FRESH)	0.63	0.32	**0.81**	0.91	**0.94**
	Discourse-aware IEP	**0.64**	**0.34**	0.71	0.8	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	0.6	**0.65**	—
	Attn.-based (FRESH)	0.63	0.52	**0.61**	0.59	**0.68**
	Discourse-aware IEP	**0.69**	**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 (FRESH)	0.53	0.52	0.71	0.71	0.82
	Discourse-aware IEP	**0.72**	0.63	**0.75**	0.67	0.81

Example of the Explanation 1

Passage

Nathan read the package of words in silence, his only motions the steady progress of his eyes and occasional replacement of pages. Allan sat nervously across from him in a chair Nathan had probably upholstered himself, a patchwork design of fabric containing easily more stuffing than any other furniture item of the period. At long last, Nathan reached the end and set down his reading on the table between them. Allan leaned forward unconsciously. "It's the best story you've ever written." Allan exhaled and leaned back into the chair, his face relaxing in imitation of his thoughts. "So," he asked, "you don't **think it's a waste of ink and paper, a futile expedition into morbidity or literary debauchery**, because I do sometimes." "Heavens, no," said Nathan, aghast. "This is one of the strongest works I've read in ages."

Question

Why Allan was nervous?

Example of the Explanation 1

Discourse-aware model

Passage

Nathan read the package of words in silence, his only motions the steady progress of his eyes and occasional replacement of pages. Allan sat nervously across from him in a chair Nathan had probably upholstered himself, a patchwork design of fabric containing easily more stuffing than any other furniture item of the period. At long last, Nathan reached the end and set down his reading on the table between them. Allan leaned forward unconsciously. "It's the best story you've ever written." Allan exhaled and leaned back into the chair, his face relaxing in imitation of his thoughts. "So," he asked, " you don't **think it's a waste of ink and paper, a futile expedition into morbidity or literary debauchery , because I do sometimes.**" "Heavens, no," said Nathan, aghast. "This is one of the strongest works I've read in ages."

Question

Why Allan was nervous?

Example of the Explanation 1

FRESH model

Passage
Nathan read the package of words in silence, his only motions the steady progress of his eyes and occasional replacement of pages. Allan sat nervously across from him in a chair Nathan had probably upholstered himself, a patchwork design of fabric containing easily more stuffing than any other furniture item of the period. At long last, Nathan reached the end and set down his reading on the table between them. Allan leaned forward unconsciously. "It's the best story you've ever written." Allan exhaled and leaned back into the chair, his face relaxing in imitation of his thoughts. "So," he asked, "you don't **think it's a waste of ink and paper, a futile expedition into morbidity or literary debauchery**, because I do sometimes." "Heavens, no," said Nathan, aghast. "This is one of the strongest works I've read in ages."
Question
Why Allan was nervous?

Example of the Explanation 2

The Stanford Sentiment Treebank

Passage

Maybe not a classic but a movie the kids will want to see over and over again.

Class: *positive*

Discourse-aware model:

Maybe not a classic but a movie the kids will want to see over and over again.

FRESH model:

Maybe not a classic but a movie the kids will want to see over and over again.

Example of the Explanation 3

The Stanford Sentiment Treebank

Passage
Marvelously entertaining and deliriously joyous documentary.

Class: *positive*

Discourse-aware model:
Marvelously entertaining and deliriously joyous documentary .

FRESH model:
Marvelously entertaining and deliriously joyous documentary.

Part 3. Summary

1. An explanation method for rationales extraction is proposed.
2. The obtained rationales are built based on the informative discourse features and are more grammatically consistent in comparison to the existing RE approaches.

Conclusion

- The impact of discourse features on the complex NLP tasks (argumentation mining, MRC, models' explanation).
- The novel methods for discourse structure encoding into pre-trained LMs have been proposed.
- An approach to retrieve explanation rationales for LMs has been presented.
- Experimental evaluation of the proposed method is performed on the English benchmarks and compared to the existing models' performance.
- We applied the disBERT model for argumentation classification of users' reviews in e-commerce chatbot[7].

[7] The code is available on github: https://github.com/lizagonch/Chatbot

References

G. Jawahar, B. Sagot, D. SeddahB.
What does BERT learn about the structure of language?. 2020.

J. Hewitt, C.D. Manning.
A structural probe for finding syntax in word representations.
NAACL HLT 2019, 2019.

D./ Yenicelik, F./ Schmidt, Y./ Kilcher.
How does BERT capture semantics? A closer look at polysemous words.
ACL 2019, 2019.

S./ Lv, D./ Guo, J./ Xu et al.
Graph-based reasoning over heterogeneous external knowledge for commonsense question answering.
AAAI 2020, 2020.

S./ Wang, J./ Zhang, C./ Zong
Learning sentence representation with guidance of human attention.
IJCAI, 2017.

References

W.C./ Mann, S.A./ Thompson
Rhetorical Structure Theory: Toward a functional theory of text organization.
Text, 1988.

S./ Jain, S./ Wiegreffe, Y./ Pinter, B.C./ Wallace
Learning to Faithfully Rationalize by Construction. 2020.