

Neural Discourse Approaches for Text Categorization and Ranking

Alexander Chernyavskiy and Dmitry Ilvovsky

National Research University Higher School of Economics
Russia, Moscow

alschernyavskiy@gmail.com, dilvovsky@hse.ru

Overview

1. RSTRecNN for single text categorization
2. DSNDM for text pairs classification and ranking
3. High-level tool for fact-checking and argument mining based on deep learning models (DEMO)
4. Further improvements to the RSTRecNN model based on additional linguistic features

RSTRecNN for single text categorization



Alexander Chernyavskiy and Dmitry Ilvovsky: “Recursive Neural Text Classification using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks”.

In: ISMIS 2020. Lecture Notes in Computer Science, vol 12117. Springer, Cham. https://doi.org/10.1007/978-3-030-59491-6_9 (2020)

Subject Area

Text classification can be applied in many industries and various applications, e.g., for filtering (including spam), personalizing advertisements, genre classification.

We consider challenging tasks in which input texts are quite long and semantic embeddings alone may not be sufficient to achieve high quality.

Some tasks require the analysis of texts in terms of its style, discourse structure.

Tasks

1) Automatic verification of factual texts

- Task: classification, labels depend on truthfulness
- Input: political statements

Nowadays, social media users are inundated with factual texts about politics, economics, history and so on. Triggered by the fact that some sources can utilize fake statements for their purposes, it would be unwise to trust all of them.

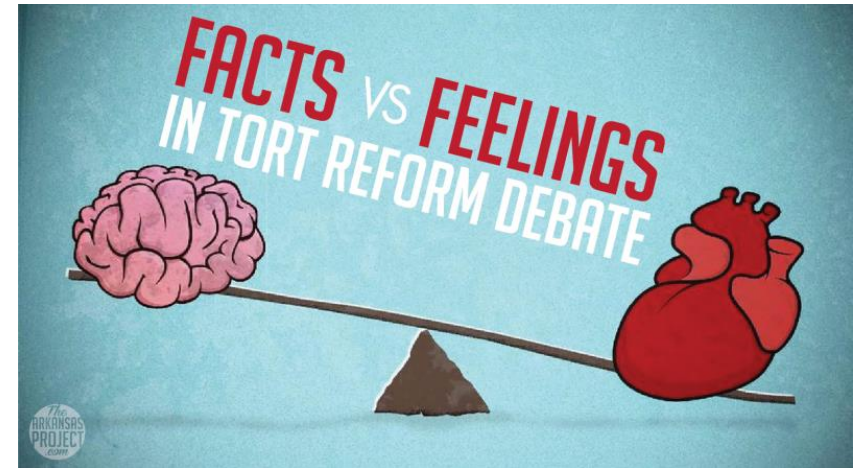


Tasks

2) Argument mining

- Task: binary classification, main goal of which is to categorize texts as either factually or emotionally justified
- Input: texts from an internet forum

This is the task of analyzing texts from the point of view of psychology and rhetoric. It can be used in dialogue systems, sentiment analysis and so on.

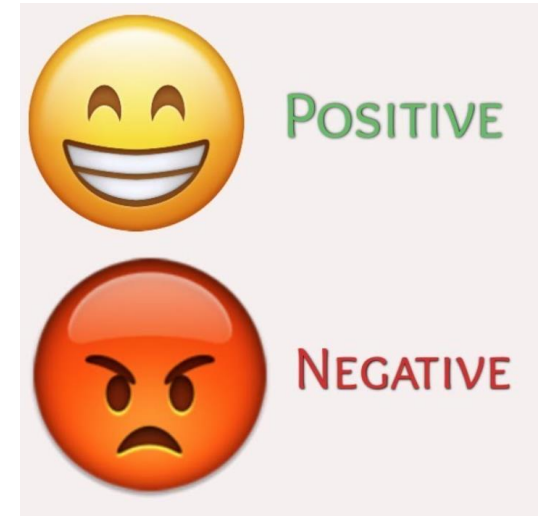


Tasks

3) Sentiment analysis

- Task: classification, main goal is to determine whether the emotional tone (sentiment) of the text is positive or negative
- Input: social-media texts

This task has become more popular over the years. It can be applied in e-commerce, marketing and advertising.



Current Approaches

1. Sequence-based approaches (usually BERT-like models)
2. Knowledge-based approaches. Such approaches can be applied in the fact-checking problem. For instance, relations between the entities of the graph can be used to confirm some “distill” information extracted from the statement. However, it is generally applicable only to factoid statements since the entities must exist within knowledge graphs.
3. Style-based approaches. These methods consider discourse text structures and make prediction using non-neural baselines like TreeKernel SVM or neural models.

Contributions

Aim: to develop a universal model capable of effectively solving the proposed tasks using discourse tree structure and TreeLSTM.

To this end, the TreeLSTM model is modified to get the universal and effective model based on the discourse tree structure.

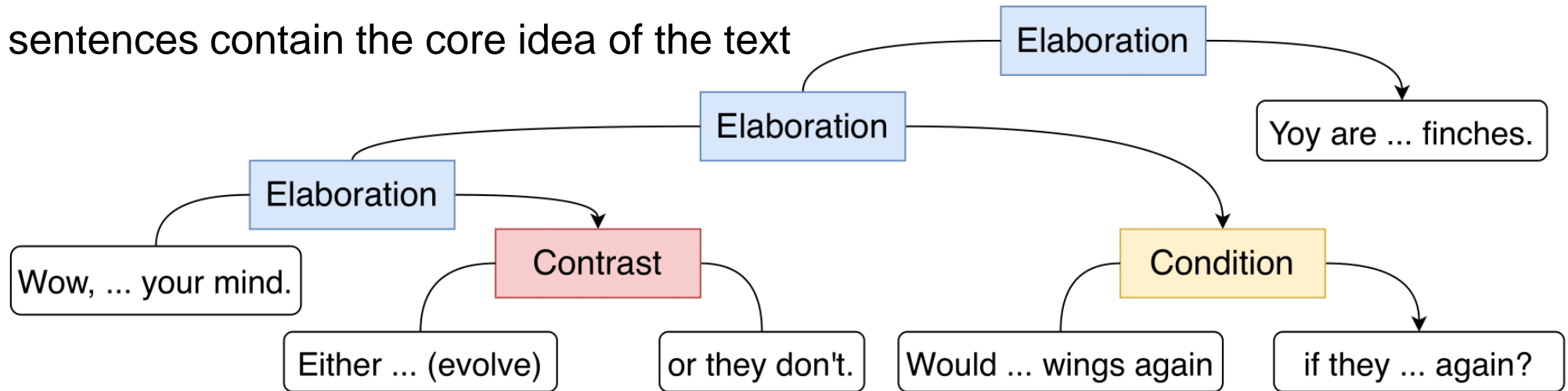
The analogous model was utilized only for unlabeled syntax trees in the sentiment analysis and semantic relatedness of two sentences task before and achieved top results.

Discourse Tree Structure

- ❖ RST posits that every document can be represented as a binary tree
- ❖ Leaves are elementary discourse units (EDUs)
- ❖ Text spans are connected recursively by discourse relations
- ❖ There two types of vertices: “Nucleus” (crucial) and “Satellite” (supplementary)
- ❖ ALT Document-level Discourse Parser can be used to construct such trees

Discourse Tree Structure

- ❖ Example of the discourse tree constructed by the ALT parser
- ❖ Here, we have 6 EDUs connected by 5 discourse relations
- ❖ Arrows are drawn from Nucleus nodes to Satellite nodes
- ❖ First sentences contain the core idea of the text



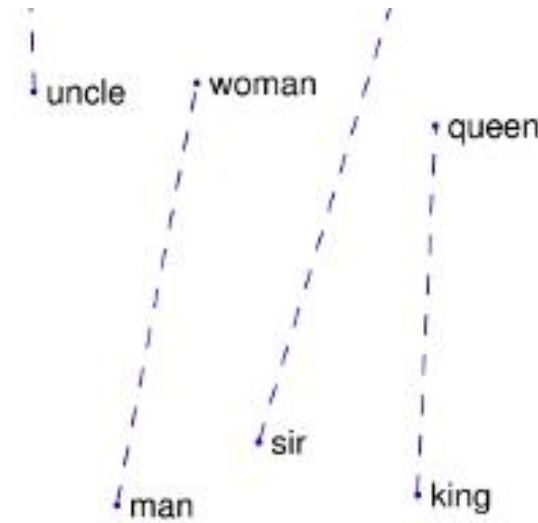
Text: "Wow, make up your mind. Either populations change over time (evolve) or they don't. Would your wingless beetles be able to produce wings again if they were somehow beneficial again? You are starting to sound like Darwin and his finches."

EDU Embeddings

1) Mean value of GloVe embeddings

This is the standard baseline word embeddings that do not consider context.

We utilized 300-dimensional pre-trained GloVe vectors.



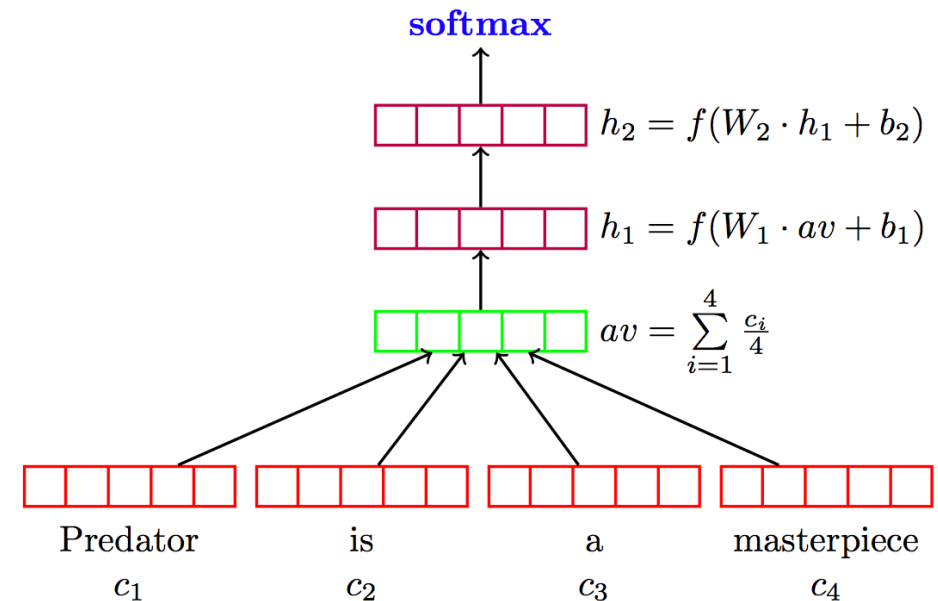
EDU Embeddings

2) Semantic embeddings from DAN.

Universal Sentence Encoder is the open-source model proposed by Google. It achieves the best results in different transfer learning tasks.

There are two modifications of its architecture: Transformer network and Deep Averaging Network (DAN), where the former is more qualitative, and the latter is more memory- and time-efficient.

DAN



EDU Embeddings

2) Semantic embeddings from DAN

One benefit of this model to EDU embeddings is that it ensures that the semantics of the context will be considered, not just the word meaning.

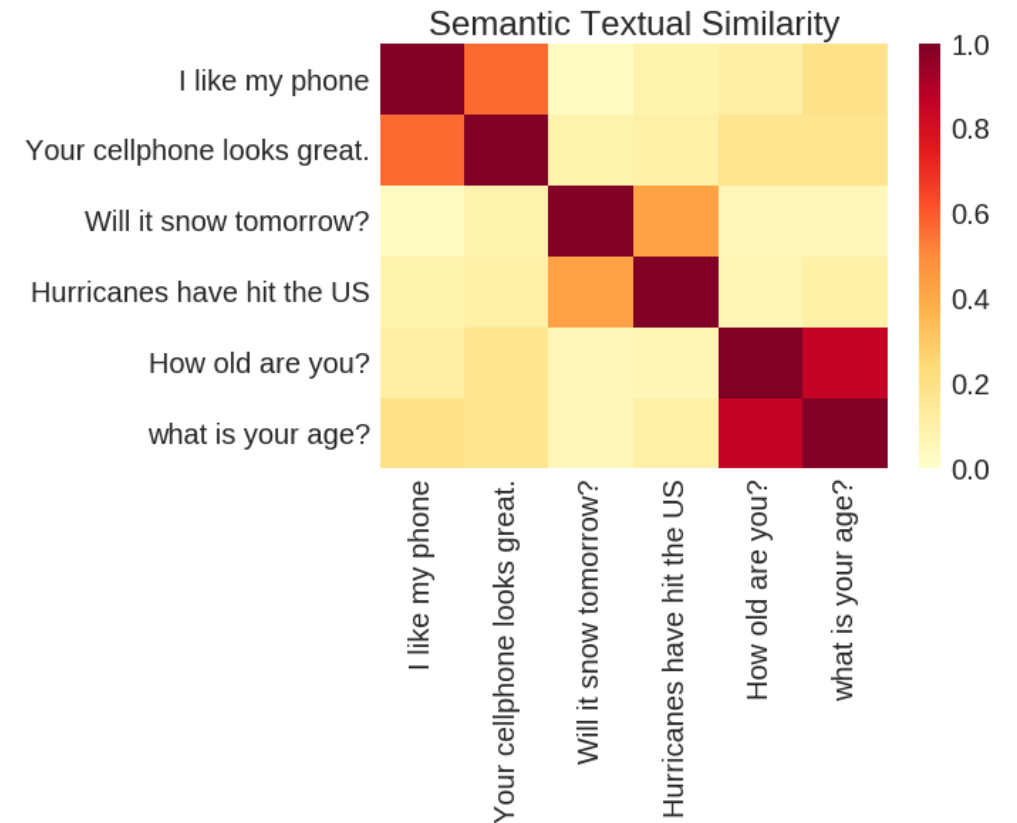
Final sentence embeddings:

DAN embeddings



concatenation

One-hot encoded POS-tags



Recursive Neural Network

Text embedding corresponding to the node i :

$$x_i = \begin{cases} \text{EDU embedding,} & \text{if } i \text{ is a leaf} \\ \text{Embedding of the empty text,} & \text{if } i \text{ is the inner node} \end{cases}$$

Text Encoder embedding:

$$\text{Text_Enc}(i) = \text{FC}(x_i)$$

Let nodes denoted as j and k be children indices for the node i , and r be the name of the discourse relation between them.

Node embedding:

$$t_i = \text{Concat}[\mathbb{I}[j \text{ is Nucleus}], \mathbb{I}[k \text{ is Nucleus}], \text{OneHot}(r), \text{Text_Enc}(i)]$$

Recursive Neural Network

Embedding of the subtree with root in the node i :

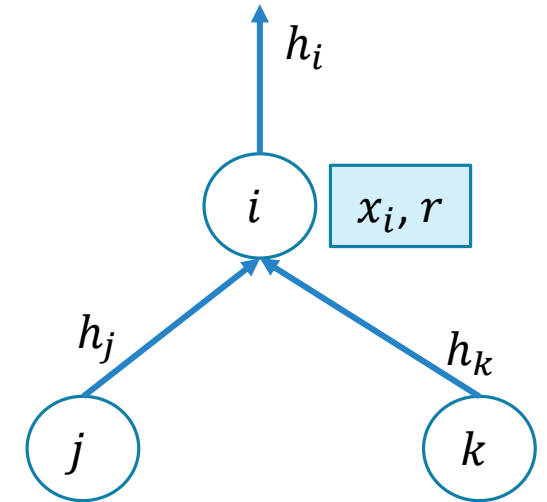
$$h_i = \text{TreeLSTM}(t_i, h_j, h_k)$$

In TreeLSTM the **memory cell** c_i is calculated as:

$$c_i = c_j * \mathbf{f}_{i_0} + c_k * \mathbf{f}_{i_1} + \mathbf{i}_i * \mathbf{u}_i$$

forget gates to
choose information
from children

input gate
values to update
with dropout inside



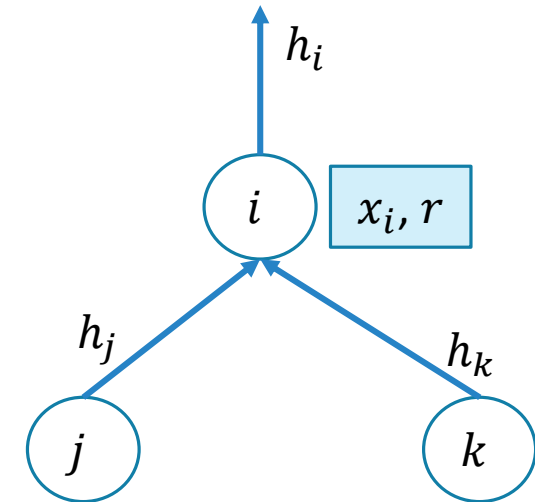
Recursive Neural Network

Full TreeLSTM equations:

$$\begin{pmatrix} \mathbf{i}_i \\ \mathbf{f}_{i0} \\ \mathbf{f}_{i1} \\ \mathbf{o}_i \\ \mathbf{u}_i \end{pmatrix} = \begin{pmatrix} \sigma(W_i[t_i, h_j, h_k] + b_i) \\ \sigma(W_{f_0}[t_i, h_j, h_k] + b_{f_0}) \\ \sigma(W_{f_1}[t_i, h_j, h_k] + b_{f_1}) \\ \sigma(W_o[t_i, h_j, h_k] + b_o) \\ D(\tanh(W_u[t_i, h_j, h_k] + b_u), \alpha) \end{pmatrix}$$

$$c_i = c_j * \mathbf{f}_{i0} + c_k * \mathbf{f}_{i1} + \mathbf{i}_i * \mathbf{u}_i$$

$$h_i = \mathbf{o}_i * c_i$$



Here, σ is the sigmoid function, D is the Dropout function and $*$ is the element-wise multiplication.

Implementation Details

- ❖ We utilized ALT Document-level Discourse Parser. Its output was transformed into the RST tree's format.
- ❖ We utilized the python Dynet library to implement the recursive neural network.
- ❖ The size of the hidden layer in LSTM cells was established at 300, the dropout rate at 0.1, the learning rate at 0.001 and the number of units in the fully-connected layer in the Text Encoder at the dimension of x_i .
- ❖ We used Dropout regularization in TreeLSTM.
- ❖ We chose the Adagrad optimizer which is less prone to overfitting for the considered tasks.

Internet Argumentation Corpus

- Messages from the Internet forum 4forums.com
- Labels: “factual” and “feeling”
- Size: 5,848 instances which were split into train, validation and test sets in the ratio 4:2:1
- Among the texts there are many voluminous ones.

IAC Experiments

The parser identified 18 unique discourse relations. It selects “Elaboration” by default. “Attribution”, “Joint” and “Same-Unit” are also popular.

“Attribution” usually corresponds to the introductory phrases like “I think that” and “I suppose that”. “Joint” is indicating the frequent use of the unions. “Same-Unit” indicates that the main meaning of the text in both sub-trees is the same.

IAC Experiments

Users employ “Elaboration” and “Same-Unit” relations more often for the argumentation based on facts and “Attribution” and “Condition” relations to express their feelings.

Discourse relation	For feeling	For factual
Elaboration	0.449	0.483
Attribution	0.132	0.107
Same-Unit	0.078	0.087
Condition	0.031	0.023

The most different relative frequencies of the relations

IAC Experiments

Модель	Признаки	Prec (fact.)	Rec (fact.)	Prec (feel.)	Rec (feel.)	Macro avg. F1
Random baseline	-	59.08	59.08	40.59	40.59	49.83
Oraby et al. (2015)	patterns	79.9	40.1	63.0	19.2	41.4
Naïve Bayes	unigrams, binary	73.0	67.0	57.0	65.0	65.0
SVM	unigrams	76.14	74.86	64.31	65.81	70.24
CNN	word2vec	82.58	84.72	76.96	74.06	79.56
	dep. embed.	78.49	77.81	68.18	69.04	73.38
	fact. embed.	76.24	74.93	64.49	66.12	70.43
	all embed.	81.98	81.27	73.14	74.06	77.61
LSTM	word2vec	80.60	77.81	69.32	72.80	75.10
	dep. embed.	78.70	76.66	67.34	69.87	73.12
	fact. embed.	78.77	81.27	71.49	68.20	74.90
	all embed.	77.09	82.42	71.63	64.43	73.75
BERT	-	84.64	80.98	74.02	76.27	79.51
Recursive model	GloVe	81.61	81.84	73.53	73.22	77.55
	DAN	83.47	85.88	78.60	75.31	80.79

IAC Experiments

- Numerous false predictions are observed for the short texts for which discourse trees have only a few nodes.
- For big trees the quality is much higher

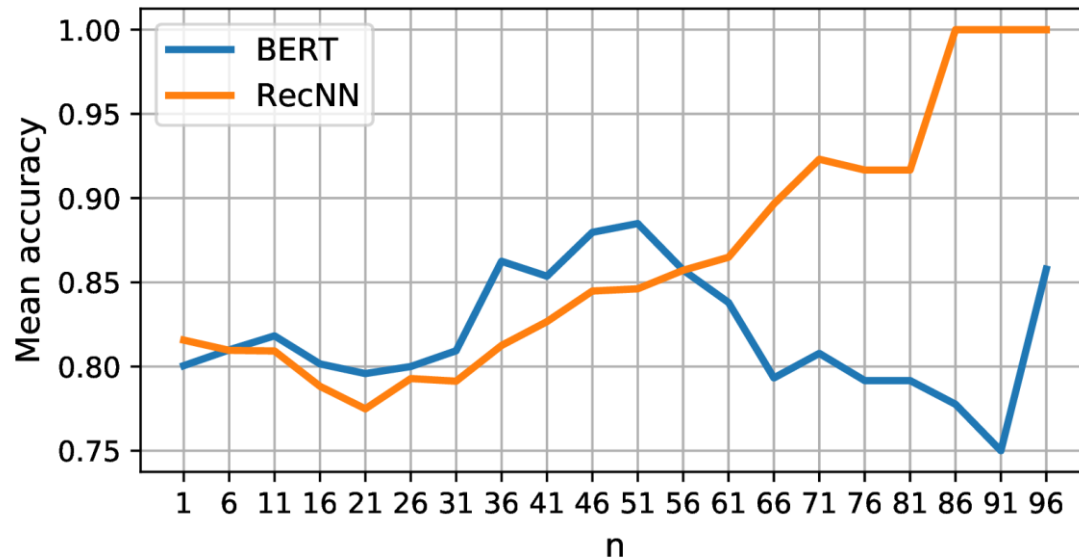
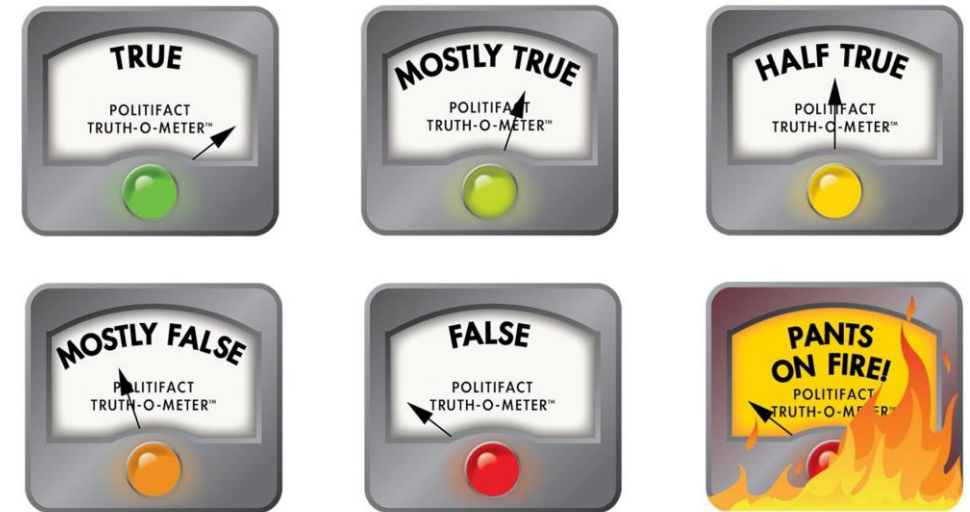


Figure demonstrates performance for the long texts. Only discourse trees that have at least n nodes are considered.

LIAR Dataset

- Data: statements collected from **politifact.com**
- Labels: 6-point scale or binary depending on truthfulness
- Size: 12,782 statements which were split into the train, validation and test samples in the ratio of 10:1:1
- The dataset is balanced, and the accuracy metric can be used to compare results.



LIAR Experiments

The discourse parser distinguished 16 unique discourse relations. The most popular relations are “Elaboration”, “Same-Unit” and “Attribution”.

Speakers use “Elaboration” and “Contrast” relations more often for the truthful statements whereas “Attribution” and “Enablement” relations are more popular for persuasion in some misleading information.

Discourse relation	“pants-fire”	“true”
Elaboration	0.377	0.412
Attribution	0.158	0.121
Enablement	0.060	0.043
Contrast	0.015	0.031

LIAR Experiments

The recursive model with DAN embeddings in EDUs got the highest results. However, the quality is limited by short claims that have trivial discourse structure.

Model	Accuracy
Majority Baseline	0.208
SVMs	0.255
Logistic Regression	0.247
LSTM + Attention	0.255
Bi-LSTMs	0.233
CNNs	0.270
CNN + LSTM	0.291
RSTRecNN (GloVe)	0.276
RSTRecNN (DAN)	0.302

Movies Dataset

- Data: movie reviews from the corpus constructed by Pang and Lee
- Labels: binary
- Size: 1000 instances for each class

The parser identified 18 unique discourse relations. The most popular relations are “Elaboration”, “Same-Unit” and “Joint”. It indicates that humans used mainly multi-nuclear relations to give their opinion about movies. Distributions of the relations are not significantly different for the given classes.

Movies Experiments

The closest work to our research is the paper by Ji and Smith. In this paper, the recursive neural network based on discourse is also proposed but has another structure.

Model	Accuracy
Hogenboom et al.	0.719
Bhatia et al.	0.829
Ji and Smith	0.831
RSTRecNN (GloVe)	0.845
RSTRecNN (DAN)	0.852

Our model outperformed other methods, and DAN embeddings availed to obtain the best quality.

Summary

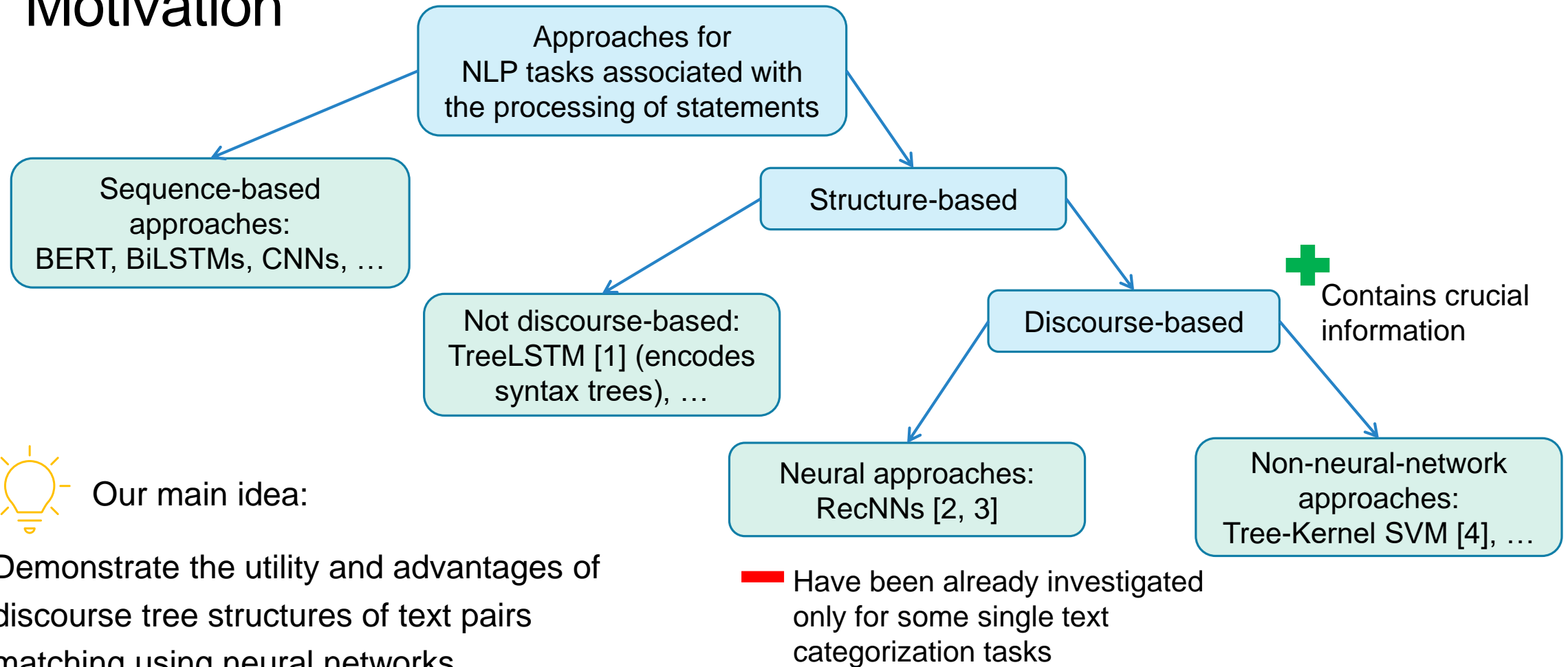
- We proposed an efficient model for text categorization using the discourse tree structure
- To this end, we modified TreeLSTM
- Experiments illustrate that the proposed models are effective and reach the best results in the assigned tasks
- The evaluation also demonstrates that discourse analysis improves quality for the classification of longer texts

DSNDM for text pairs classification and ranking



Alexander Chernyavskiy and Dmitry Ilvovsky: “DSNDM: Deep Siamese Neural Discourse Model with Attention for Text Pairs Categorization and Ranking”. In: Proceedings of the First Workshop on Computational Approaches to Discourse at EMNLP 2020, pp. 76-85, (2020)

Motivation



Our main idea:

Demonstrate the utility and advantages of discourse tree structures of text pairs matching using neural networks

Tasks

Main criterion: input texts are quite long, and paragraphs are given initially.

1) Automatic verification of factual texts

- Task: classification
- Input: statements and their justifications or refutations

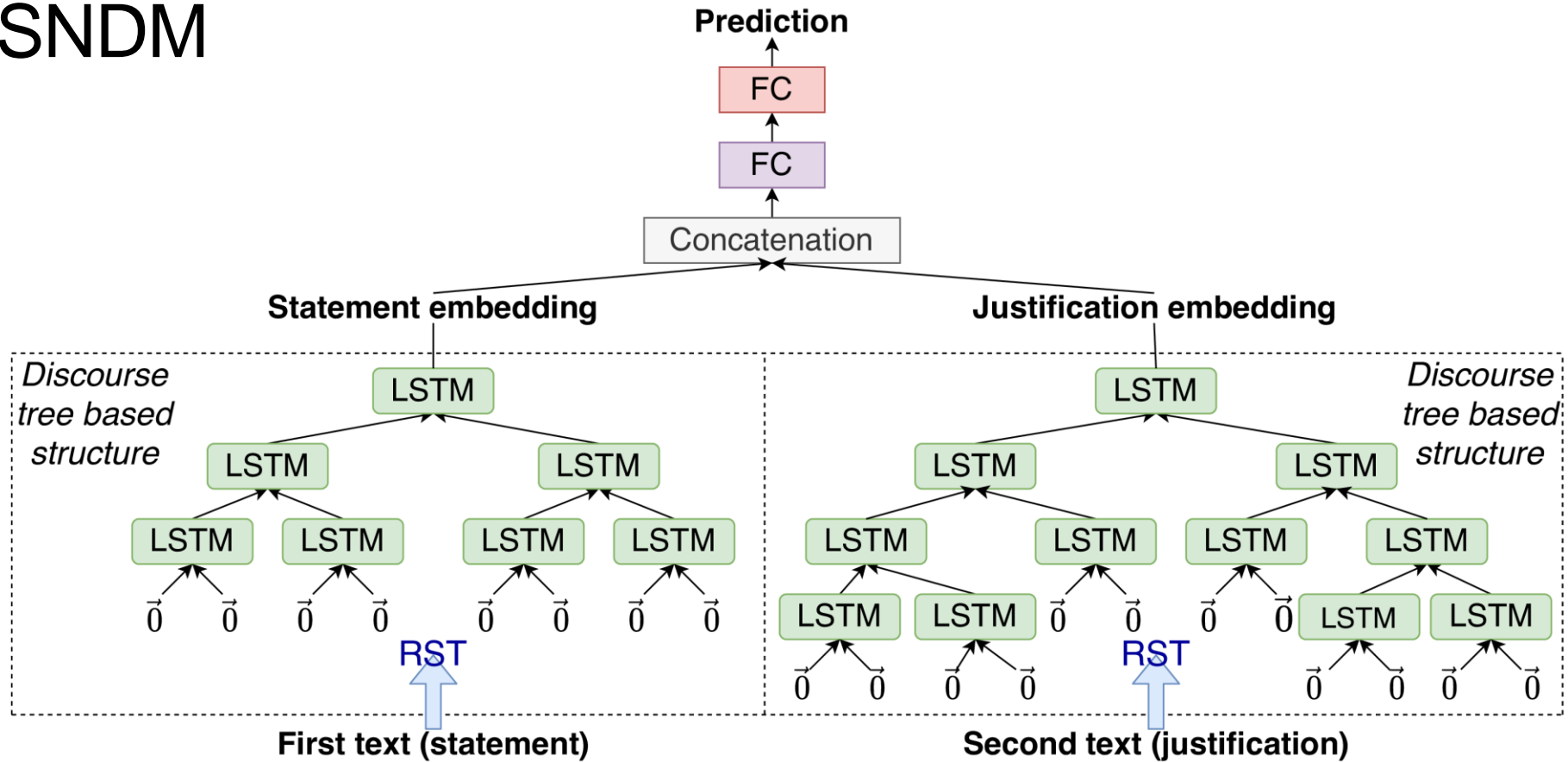
2) Question answering systems

- Task: ranking
- Input: non-factoid questions with possible set of answers

Contributions

- We propose DSNDM (Deep Siamese Neural Discourse Model) to analyze pairs of texts.
- We suggest the way of integration of the attention mechanism in the DSNDM model.
- We investigate the value of the proposed approach considering two tasks and experimentally confirm the utility and importance of discourse analysis for the text pairs processing.

DSNDM



“LSTM” applies the TreeLSTM cell. Cells with the same color use the same weights. Each cell applied to EDUs receives zero vectors as embeddings of its children.

Attention Mechanism

Idea: let's use a question/statement embedding to filter information while constructing an answer/justification embedding. At each step, the model decides information from which subtree is more useful.

Attention module with

- ❖ key: k is a question/statement embedding
- ❖ values = queries: $Q = [q_1, q_2] := [c_j * \mathbf{f}_{i_0}, c_k * \mathbf{f}_{i_1}]$

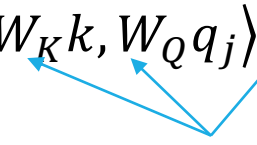
Attention Mechanism

Then, the memory cell: $c_i = 2 \cdot \text{Att}(k, Q) + \mathbf{i}_i * \mathbf{u}_i$ **(1)**

instead of $c_i = q_1 + q_2 + \mathbf{i}_i * \mathbf{u}_i$ **(2)** from RSTRecNN

$$\text{Att}(k, Q) = \text{Softmax} \left(\sum_{j=1}^{|Q|} \langle W_K k, W_Q q_j \rangle W_V q_j \right)$$

Trainable matrices of parameters



Here, Softmax layer which is used for normalization. Multiplication in **(1)** by 2 is necessary to maintain a balance with **(2)**.

Equation **(1)** is utilized instead of **(2)** only to construct the embedding of the second text (we should know k).

Training Techniques for Ranking

1. Classification based

- The ranking of the answers for each question is carried out using the class probabilities predicted by the model

2. Pointwise ranking

- Regression for pairs $\{(q_i, a_i)_{i=1..N}\}$ with relevance scores $\{r_i\}$
- Loss: $\sum_{i=1}^N (\text{DSNDM}(q_i, a_i, w) - r_i)^2 \rightarrow \min$

3. Pairwise ranking

- Inputs are triplets $\{(q_i, a_i^+, a_i^-)_{i=1..M}\}$ with the relevant and irrelevant answers
- Loss: $\sum_{i=1}^M \frac{1}{1 + \exp(\text{DSNDM}(q_i, a_i^+, w) - \text{DSNDM}(q_i, a_i^-, w))} \rightarrow \min$

LIAR-PLUS Dataset

- Data: statements from LIAR; additionally, contains justification for each statement
- The LIAR-PLUS dataset can be used in four scenarios, depending on the restriction on the available data: S (only statement is used), S + M (statement and metadata), SJ (pairs: statement and justification), and S + JM (all available data). The model proposed in this paper is applied to pairs in the SJ scenario.



LIAR-PLUS Experiments

- ❑ DSNDM model significantly improves the results of baselines, especially in the case of the multiclass classification.
- ❑ The fully-connected layer in the Text Encoder is crucial since it adds up to 0.02 to accuracy.
- ❑ The DSNDM model with the integrated attention module reached the best results for the test set. This improvement is not significant because of the binary structure of trees (the attention module re-weights only two vectors at each node).

Model	Binary		Six-way	
	valid	test	valid	test
LR	0.68	0.67	0.37	0.37
SVM	0.65	0.66	0.34	0.34
BiLSTM	0.70	0.68	0.34	0.31
P-BiLSTM	0.69	0.67	0.36	0.35
DSNDM	0.71	0.69	0.40	0.40
DSNDM + Att.	0.70	0.71	0.40	0.41

Results on the test set

Error Analysis

- Numerous false predictions are observed for the short texts for which discourse trees have only a few nodes
- For the deepest trees which contain more than 45 nodes in the statement and justification in total, the F1-score metric is higher than 0.46

True label	pants-fire	29	31	18	6	2	6
	false	20	114	31	34	11	40
	mostly false	13	43	78	41	24	15
	half-true	5	52	31	93	51	35
	mostly-true	2	42	9	52	95	49
	true	10	29	11	13	30	118
		pants-fire	false	mostly false	half-true	mostly-true	true
		Predicted label					
		Confusion matrix					

Error Analysis

Typical examples when the model predicts a wrong label:

- Refutation partly repeats the statement. Then, the model with attention focuses mainly on the repeated part and marks the misleading statement as “true”.
- Justification text is extracted inaccurately and is not sufficient to estimate the veracity of the statement.
- Justification text is complex and contain only one useful sentence.
- Justification indicates that the statement can be labeled in the opposite way in some general cases.

Error Analysis

The quality of the proposed model is limited by several factors:

- the size of the discourse trees
- the quality of the discourse parser
- the quality of the provided justifications

ANTIQUUE Dataset

- Factoid questions selected from the Yahoo! Webscope L6 database
- In total: 2,426 questions and 27,422 answers in the training sample
200 questions and 6,589 answers in the test sample
- Labels: 4-point scale depending on the relevance of answers
- Questions contain ~11 words and answers ~47 words on average



ANTIQUÉ Experiments

- The discourse parser identified 18 different discourse relations
- Ranking metrics:

Let $\{q_j\}_{j=1\dots N}$ is the set of questions and $I(q)$ is a sorted set of relevant answers for each question. Let the model predict a ranked list of answers $J(q)$ for each question, and $J_k(q)$ are its first k elements (the model considers them the most relevant).

$$MRR = \frac{1}{N} \sum_{j=1}^N \frac{1}{\text{position of the first relevant answer in } J(q_j)}$$

$$P@K = \frac{1}{N} \sum_{j=1}^N \frac{|J_K(q_j) \cap I(q_j)|}{\min(|I(q_j)|, K)}$$

ANTIQUÉ Experiments

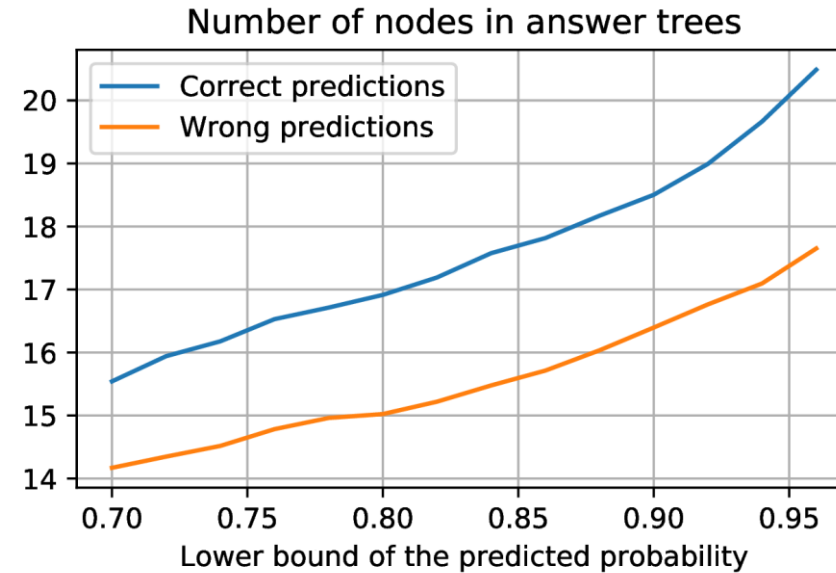
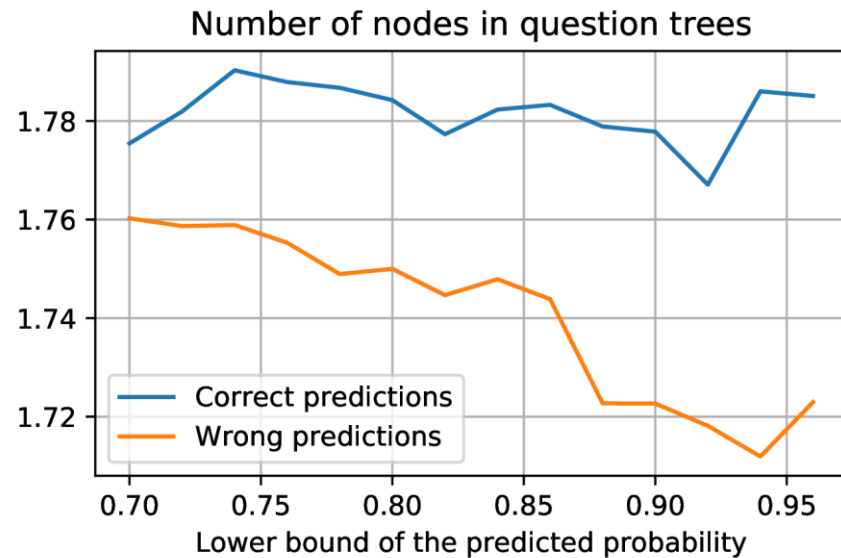
Comparison with:

- Models presented in [7]. The results were obtained on the **extended dataset**, and it is hard to reproduce them.
- Models presented in [8]. We do not compare our results with the results obtained with a modified curriculum since we consider only the basic pointwise and pairwise losses.

	Model	MRR	P@1
1	BM25	0.4885	0.3333
	DRMM-TKS (2016)	0.5774	0.4337
	aNMM (2016)	0.6250	0.4847
	BERT (2018)	0.7968	0.7092
2	ConvKNRM [pairwise]	0.4920	0.3650
	BERT [pointwise]	0.6694	0.5550
	BERT [pairwise]	0.6999	0.5850
	Tuned BM25	0.5802	0.4550
	Tuned SDM	0.5377	0.4400
3	Base [classif.]	0.6792	0.5350
	+ Att. [classif.]	0.6830	0.5350
	Base [pointwise]	0.6864	0.5300
	+ Att. [pointwise]	0.7098	0.5650
	Base [pairwise]	0.7120	0.5800
	+ Att. [pairwise]	0.7267	0.6000

Results on the test set. 1: Models presented in [7], 2: Models presented in [8], 3: DSNDM

Error Analysis



- a) number of nodes in correct cases is greater than number of nodes in wrong cases
- b) model's greater confidence in the wrong answer is frequently triggered by the smaller size of the question tree

Dependence of the number of nodes in the discourse trees of questions and answers on the confidence of DSNDM in cases of correct and wrong predictions

Error Analysis

Thus,

1. for both questions and answers, the number of nodes in discourse trees for correctly classified pairs is greater than for incorrectly classified ones. Therefore, DSNDM makes wrong predictions mostly for small trees.
2. model's greater confidence in the wrong answer is frequently triggered by the smaller size of the question tree.

Error Analysis

Typical examples when the model predicts a wrong label:

- The question contains only a few significant keywords, and the model focuses mainly on them. Even though the answer is irrelevant and unrelated to the question area, it often uses the same keywords. Thus, similar EDU embeddings do not contribute to the correct classification.
- The meaning of the answer and the question is the opposite. That is, despite the correctness of the answer, its text refutes the information in the question.
- The correct answers may be formulated in the way not expected by the authors of the questions.

Thus, the quality of the model is also limited by the variability of possible answers.

Summary

- ❖ We investigated the utility and importance of the discourse analysis for text pairs categorization and ranking
- ❖ To this end, we proposed DSNDM which is capable of processing pairs of texts
- ❖ In addition, we proposed the integration of the attention mechanism in DSNDM
- ❖ DSNDM efficiently learned the match between discourse tree structures and achieved high quality in both considered tasks
- ❖ The error analysis demonstrates that the model processes deeper trees more successfully

Future Work

1. Applying trees not only of a binary structure
2. The modification of vector representations of EDUs
3. The investigation of the performance of DSNDM in other tasks where discourse analysis may be helpful. For instance, in machine translation, chat-bots and other QA systems.

High-level tool for fact-checking and argument mining based on deep learning models (DEMO)

Project Goals

Application capable of solving the two following tasks:

1. Automatic verification of political statements
 - a) Classification of single texts
 - b) Classification of pairs of texts (statements + justifications or refutations)
2. Analysis of the argumentation of the text in terms of the presence of a factual or emotional basis

Interpretability: the user should be able to see which parts of the text influenced the model's prediction the most.

Constituents

1. Model architecture:

- RSTRecNN for single text categorization
- DSNDM for pairs text categorization

2. EDU embeddings

- Pretrained DAN model from the Universal Sentence Encoder

3. Planning: how to interpret the results?

- The model gets trees with “hidden” EDUs as the input (the corresponding leaf contains empty text). Then, the crucial span is the EDU after the removal of which the probability of the predicted class changed the most.

Implementation Details

- Implementation in Python; model within the Dynet library (with C++ code integration)
- Training on the LIAR and LIAR-PLUS datasets for the fact-checking task, and on the Internet Argument Corpus for the argumentation analysis task
- Interface in the format of a web application with additional functionality for interpretation and trees visualization
- The application has the functionality of internal caching
- Deploy via docker container

Application Interface

Modes:

1. Political Fact-checking
2. Detection of Factual Argumentation

External settings:

1. whether to show the interpretation? (if yes, how many most important EDUs to underline -- an integer, not less than 1)
2. whether to visualize discourse trees?

Output:

1. Mandatory: the label of the predicted class
2. Optional: display interpretation and display of discourse trees

Examples of Work (Fact-checking)

Settings

Task

Fact-checking

☒ Show interpretation

☒ Draw discourse trees

Select the number of the important EDUs to highlight

2

Political fact-checking

Input

Political statement

John Holdren, director of the White House Office of Science and Technology Policy, has proposed forcing abortions and putting sterilants in the drinking water to control population.

Justification (if provided)

Check the input statement

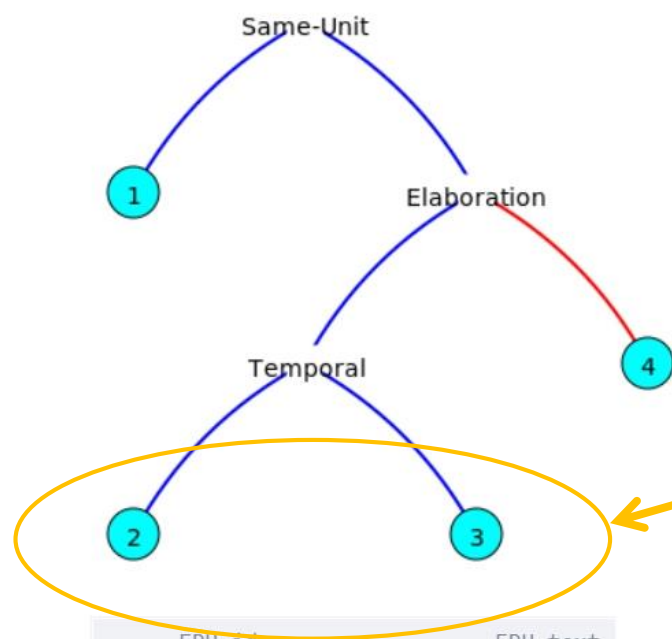
Result

The Predicted Class is: True



“John Holdren, director of the White House Office of Science and Technology Policy, has proposed forcing abortions and putting sterilants in the drinking water to control population.”

Examples of Work (Fact-checking)



Interpretation

EDU importances for the model (for the preprocessed text):

John Holdren , director of the White House Office of Science and Technology Policy , has proposed forcing abortions and putting sterilants in the drinking

0.256

0.266

water to control population .

	EDU id	EDU text
0	1	John Holdren , directo...
1	2	has proposed forcing a...
2	3	and putting sterilants...
3	4	to control population .

Blue arrows are drawn to the Nucleus nodes and red arrow are drawn to the Satellite nodes.

Justification (if provided)

Holdren's position. Holdren didn't advocate those ideas then, and, when asked at a Senate confirmation hearing, Holdren said he did not support them now. We think it's irresponsible to pluck a few lines from a 1,000-page, 30-year-old textbook, and then present them out of context to dismiss Holdren's long and distinguished career.

Check the input statement

Result

The Predicted Class is: **False**



Interpretation

EDU importances for the model (for the preprocessed text):

John Holdren, director of the White House Office of Science and Technology
0.259

Policy, has proposed forcing abortions and putting sterilants in the drinking
0.255

water to control population .

Let's add the refutation text:

“But in seeking to score points for a political argument, Beck **seriously mischaracterizes Holdren's positions**. Holdren didn't advocate those ideas then. And, when asked at a Senate confirmation hearing, Holdren said he did not support them now. We think it's irresponsible to pluck a few lines from a 1,000-page, 30-year-old textbook, and then present them out of context to dismiss Holdren's long and distinguished career.”

Examples of Work (Argument Mining)

“Such personal attacks as yours indicate a bankruptcy of argument and reveal much about you while revealing nothing about me. **If someone says** that you are more likely to die in a car accident than in an airplane accident would accuse them of not caring about airplane accident victims? Would you accuse them of taking delight in the number of car deaths?”

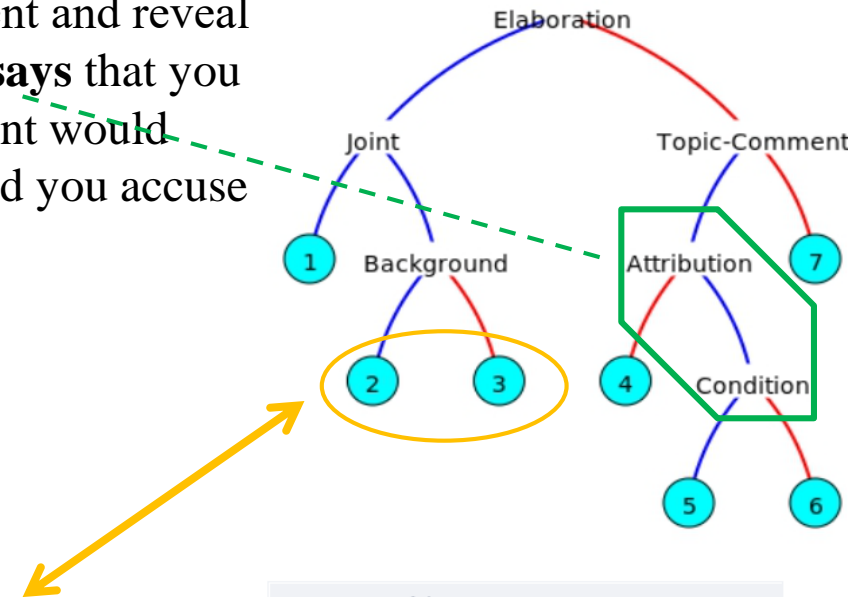
Result

The Predicted Class is: feeling

Interpretation

EDU importances for the model (for the preprocessed text):

Such personal attacks as yours indicate a bankruptcy of argument and reveal much
about you while revealing nothing about me . If someone says that you are more
likely to die in a car accident than in an airplane accident would accuse them
of not caring about airplane accident victims ? Would you accuse them of taking
delight in the number of car deaths ?



	EDU id	EDU text
0	1	Such personal attacks ...
1	2	and reveal much about ...
2	3	while revealing nothin...
3	4	If someone says
4	5	that you are more like...
5	6	than in an airplane ac...
6	7	Would you accuse them ...

How to run it yourself?

The project was deployed using a docker container and is located in the docker-hub along the following path: **alchernyavskiy/dyalt-project:v1**

Running instructions:

1. `docker run -it -p [X]:8501 alchernyavskiy/dyalt-project:v1`
2. `streamlit run streamlit_demo.py` [or `streamlit_drawer.py` for trees visualization]
3. open in browser `localhost:[X]`
4. some application logs will be available in the terminal during running

Further improvements to the RSTRecNN model based on additional linguistic features

Idea

There are some models that successfully use coreference graphs to improve span embeddings while solving text summarization task (e.g., DiscoBERT).

It has also been shown that syntactic graphs are useful in classification problems.

The RSTRecNN model use only the discourse structure with the semantic embeddings in its leaves.



The main idea to improve EDU embedding using syntactic and coreference graphs

Graph Convolutional Network

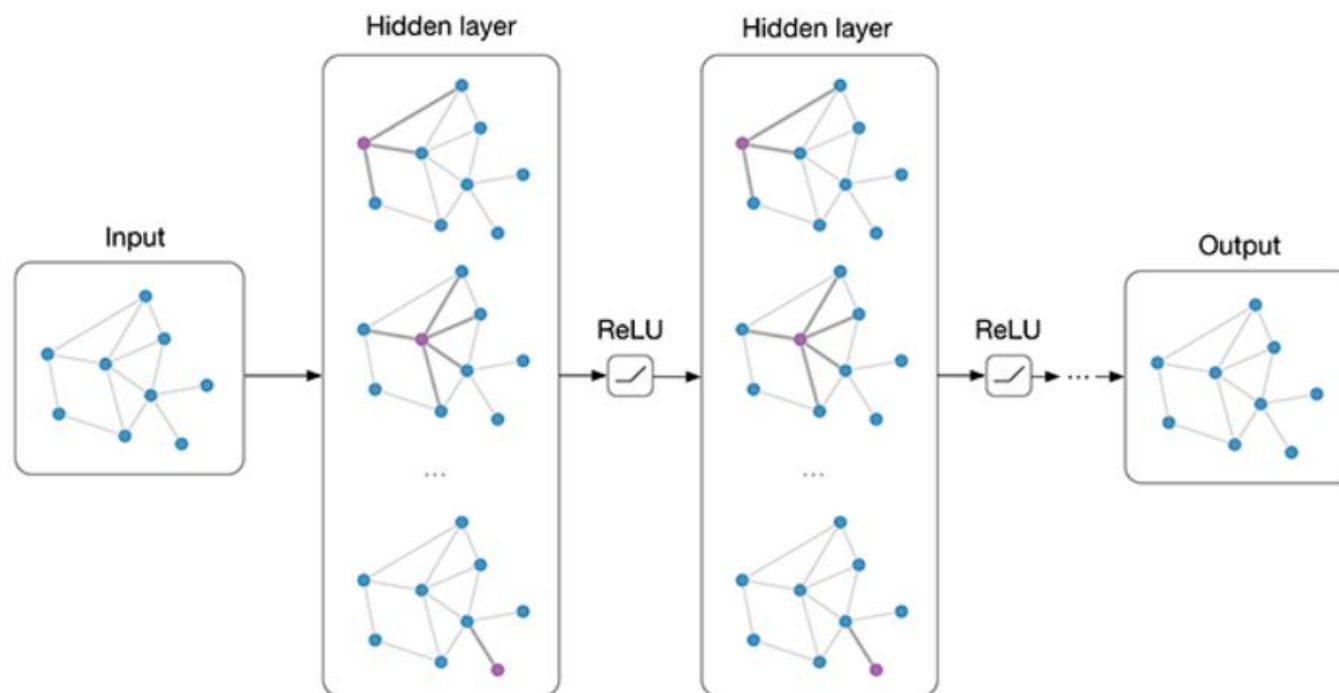
GCN allows to update nodes embeddings in any graph using its edges. The graph can be directed or undirected.

In our case, there are two options:

- 1) Vertices are words, and we can recalculate their embedding using syntactic or coreference links. Syntactic links are obtained from the dependency graph.
- 2) Vertices are EDUs, and we can recalculate their embedding using or coreference links. Here, two EDUs are connected if some words within them are connected.

Graph Convolutional Network

At each iteration, the node embedding is recalculated using current embeddings of its neighbors.



Graph Convolutional Network

Let A be the adjacency $n \times n$ matrix, and B be the embedding $n \times k$ matrix.

We should **normalize** A such that all rows sum to one:

$$\hat{A} = D^{-1}A, \text{ where } D \text{ is the diagonal node degree matrix.}$$

\hat{A} corresponds to taking the average of neighboring node features.

Input of the GCN module: $E_0 = BW_0 \tilde{+} b_0$

$W_d \in \mathbb{R}^{n \times m}$ and $b_d \in \mathbb{R}^{1 \times m}$ are trainable parameters that are needed for dimension reduction. Here, $\tilde{+}$ is column-wise summation.

Graph Convolutional Network

Equation of the i -th GCN layer:

$$E_i = E_{i-1} + \text{ReLU}(\hat{A} \cdot E_{i-1} \cdot W_i \tilde{+} b_i)$$

$W_i \in \mathbb{R}^{m \times m}$ and $b_i \in \mathbb{R}^{1 \times m}$ are trainable parameters of the GCN model.

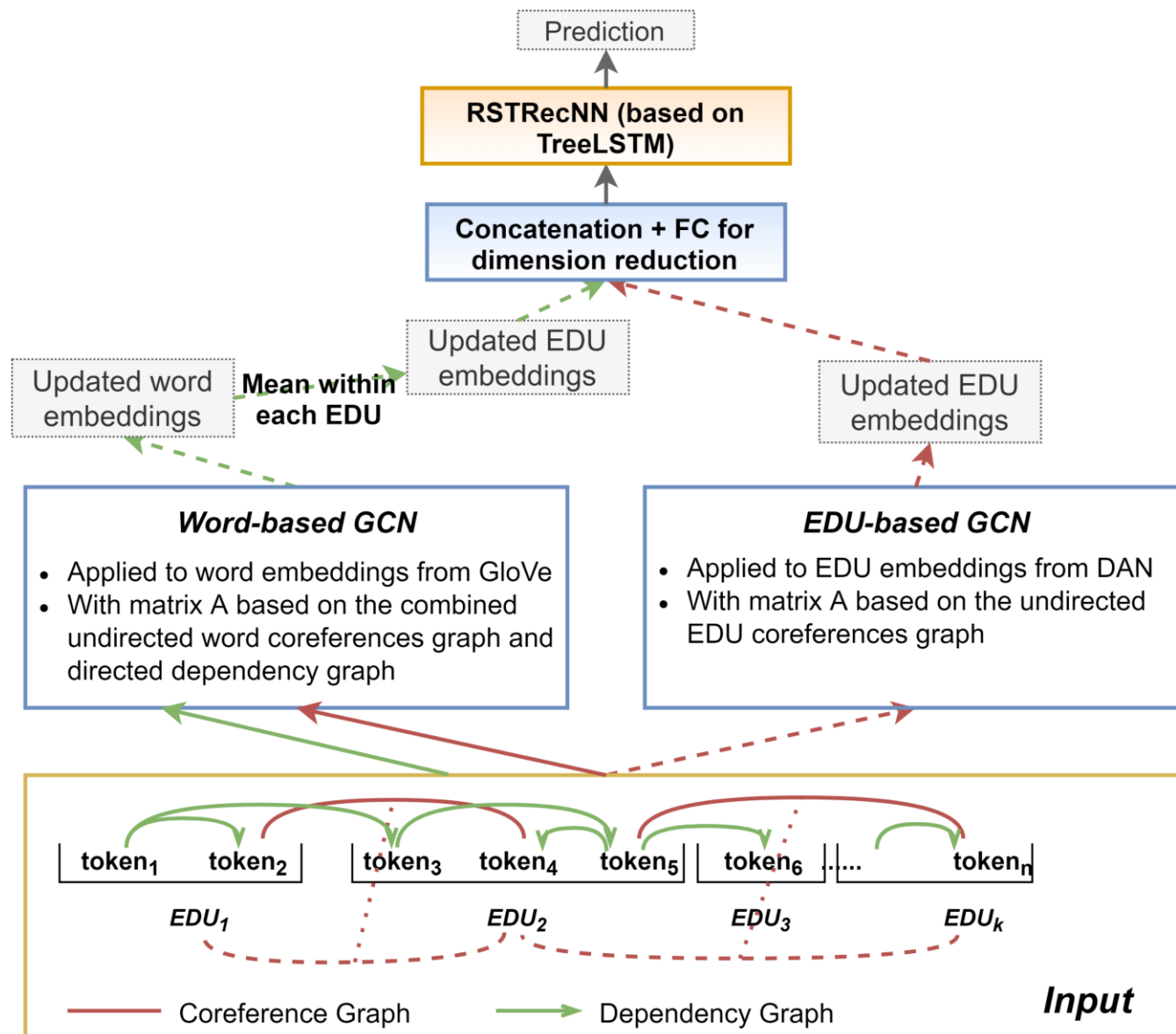
We need **residual** connection to save important information from previous steps.

Index i is iterated over the interval $[1, \dots, N]$, where N is the number of GCN layers

Output: updated embedding matrix E_N

GCN-RSTRecNN

- ✓ Two GCN modules are used in parallel.
- ✓ The first one recalculates word embeddings using a combination of coreference and dependency graphs.
- ✓ The second uses a coreference graph over the EDUs and updates the raw EDU embeddings too.
- ✓ Further, concatenation and a fully-connected layer are applied to obtain **"super"** EDU embeddings that consider **semantics, syntax, and coreferences**.



Datasets

We conducted experiments only on datasets where the texts are long (the graphs are non-trivial). That is, the LIAR dataset does not fit.

We consider **IAC** and **Movies** datasets (were used in experiments for the standard RSTRecNN model).

Also, the **Debates** dataset was used. It contains ~3000 long congressional floor-debate transcripts. The goal is to predict the vote (“yea” or “nay”) for the speaker of each speech segment.

Experiments

Configuration **a-b** means that **a** layers in the Word-based GCN, and **b** layers in the EDU-based GCN.

Config	IAC (F1-score)	Movies (Acc.)	Debates (Acc.)
0-0 (no GCN)	0.792	0.843	0.668
0-1	0.790	0.863	0.697
1-0	0.791	0.852	0.717
1-1	0.803	0.853	0.705
0-3	0.793	0.863	0.682
3-3	0.796	0.873	0.675
0-4	0.794	0.873	0.675
4-0	0.796	-	0.682
4-4	0.801	-	0.675

Results of GCN-RSTRecNN on the test sets.

There was no goal here to get the best quality. The table shows the average. The model parameters were not tuned much.

Experiments

- The best results are obtained with models that use GCN.
- It is not always obvious how many layers to take.
- The word-based GCN is more useful than the EDU-based, as it has bigger graphs.

Ablation study

We conducted experiments with random graphs to confirm the importance of the coreference and dependency graphs.

Random graphs

Config	IAC (F1-score)	Movies (Acc.)
1-0	0.785	0.832
1-1	0.796	0.863
0-3	0.782	0.852
3-3	0.794	0.863



Correct graphs

Config	IAC (F1-score)	Movies (Acc.)
1-0	0.791	0.852
1-1	0.803	0.853
0-3	0.793	0.863
3-3	0.796	0.873

In both cases, the quality for random graphs is worse.

Ablation study

We replaced RSTRecNN with a stack of fully-connected layers to check the importance of the discourse structure.

Also, we used only EDU-based GCN. This is similar to the case when there are no layers in the Word-based GCN since raw GloVe embeddings do not help to DAN embeddings.

Config	IAC (F1-score)	Movies (Acc.)	Debates (Acc.)
x-0	0.782	0.863	0.655
x-3	0.791	0.863	0.672
x-4	0.789	0.868	0.675





Conclusion:

The quality is slightly worse. Therefore, sometimes it is useful to use only the discourse segmenter and recalculate EDU embeddings.

Summary

- We proposed a way of construction the linguistic-based classification model that consider semantics, syntax, discourse and coreferences simultaneously.
- To this end, we combined GCN (based on the dependency and coreference graphs) and the RSTRecNN model.
- We investigated the importance of the used linguistic features.

References

-  [1] Kai Sheng Tai, Richard Socher, and Christopher D. Manning (2015)
Improved semantic representations from tree-structured long short-term memory networks
[CoRR, abs/1503.00075](#)
-  [2] Yangfeng Ji and Noah A. Smith (2017)
Neural Discourse Structure for Text Categorization
[Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics](#), pp. 996–1005
-  [3] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein (2015)
Better Document-level Sentiment Analysis from RST Discourse Parsing
[Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing](#), pp. 2212–2218
-  [4] Boris Galitsky, Dmitry Ilvovsky, and Sergei O. Kuznetsov (2015)
Text classification into abstract classes based on discourse structure
[In RANLP](#), pages 200–207

References

-  [5] Shafiq Joty, Giuseppe Carenini, and Raymond Ng (2012)
A Novel Discriminative Framework for Sentence-Level Discourse Analysis
[Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning](#), pp. 904-915

-  [6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil (2018)
Universal sentence encoder
[CoRR](#), [abs/1803.11175](#)

-  [7] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft (2020)
Antique: A nonfactoid question answering benchmark
[In Advances in Information Retrieval](#), pages 166–173, Cham. Springer International Publishing

-  [8] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder (2020)
Training curricula for open domain answer re-ranking
[In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval](#), pages 529–538