En-for-Motion Chatbot - Can RST relations help assess one's mental health?

Natalia Redwood
University of Colorado Boulder
Boulder, CO, United States
nawo2224@colorado.edu

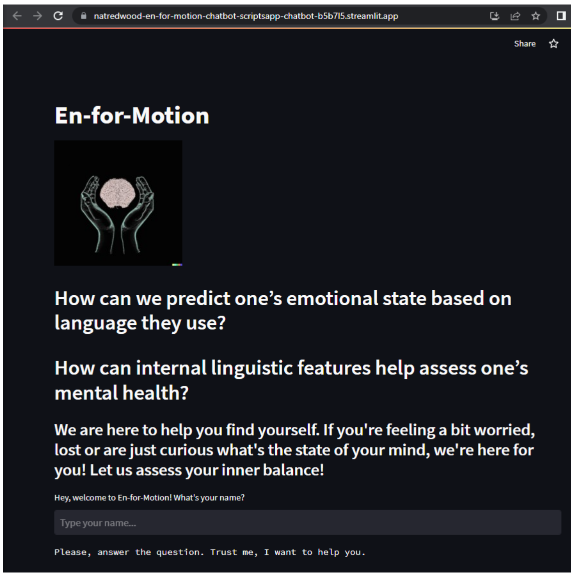
Abstract

The paper aims to present the En-for-Motion Chatbot. The project is divided into two parts. First, the author describes the chatbot's API and details of the system's model. The project attempts to check if extracting linguistic features, such as RST discourse relations, can increase the performance of emotion recognition models and therefore, help improve one's mental health using a chatbot API. The En-for-Motion Chatbot is deployed with Streamlit framework and is available for users: talk to En-for-Motion Chatbot. The code is available on GitHub (https://github.com/NatRedwood/En-for-Motion-Chatbot). Second part investigates the correlations between discourse relations and emotion labels, joy, sadness, anger and fear. 95 fairy tales (Alm, 2008) were parsed with RST text-level discourse parser (Li et al., 2014) to extract RST features per sentence. They are compared to emotion labels. The paper provides the analysis of the correlations and attempts to check if RST linguistic features can improve emotion recognition for narrative texts. To answer this question, the best performing experiemntal model is used in the En-for-Motion Chatbot API.

1 Introduction

1.1 AI and Mental Health

Artificial Intelligence (AI) has become an integral part of various industries, including healthcare. One of the areas where AI has shown promise is mental health. With the increasing prevalence of mental health disorders worldwide, there is a growing need for effective and efficient methods to diagnose and treat such disorders. Chatbots, powered by AI, have emerged as a promising tool in mental health diagnosis and therapy.


natredwood-en-for-motion-chatbot-scriptsapp-chatbot-b5b7l5.streamlit.app
Share

En-for-Motion

How can we predict one's emotional state based on language they use?

How can internal linguistic features help assess one's mental health?

We are here to help you find yourself. If you're feeling a bit worried, lost or are just curious what's the state of your mind, we're here for you! Let us assess your inner balance!

Hey, welcome to En-for-Motion! What's your name?

Type your name...

```
Please, answer the question. Trust me, I want to help you.
```

Figure 1: En-for-Motion Chatbot API deployed in Streamlit. Welcome page.

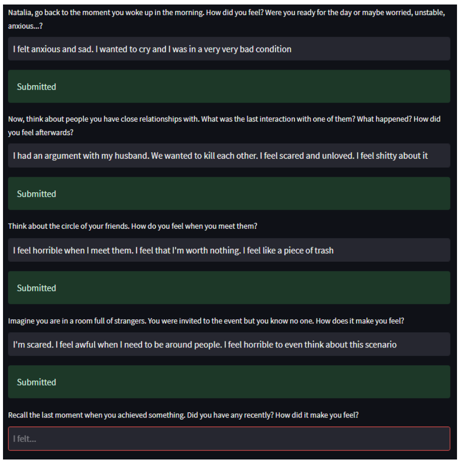
In this paper, the author introduces the En-for-Motion chatbot, a mental health chatbot that uses the Rhetorical Structure Theory (RST) to assess an individual's mental health. RST is a linguistic theory that aims to identify how text is structured to convey meaning. In this study, we hypothesize that the RST relations, when applied to chatbot conversations, can help assess an individual's mental health status accurately.

The primary aim of this paper is to investigate the potential of the En-for-Motion chatbot in assessing an individual's mental health. We will discuss the methodology used in building the chatbot and evaluating its effectiveness in assessing mental health. Furthermore, we will present the results obtained using RST supported model and traditional text classification model for Emotion Recognition (ER).

Overall, this paper aims to contribute to the ongoing research on AI and mental health by demonstrating the potential of chatbots in assessing mental health disorders. The results obtained from this study can help pave the way for the development of more accurate and efficient mental health assessment tools, possibly using linguistic features, such as RST relations.

2 En-for-Motion Chatbot API

En-for-Motion Chatbot is an NLP model based program that asks 8 questions about the mental health condition. Based on the answers, it calculates the mental health score of the user. To get the predictions, text classification model is trained on the dataset of sentences labeled for 6 emotions, joy, love, sadness, fear, anger and surprise. The labels are then put into 2 buckets: positive (joy and love) and negative (sadness, fear, anger and surprise). All 8 questions are included in the Figures 2 and 3.



Natalia, go back to the moment you woke up in the morning. How did you feel? Were you ready for the day or maybe worried, unstable, anxious...?

I felt anxious and sad. I wanted to cry and I was in a very very bad condition

Submitted

Now, think about people you have close relationships with. What was the last interaction with one of them? What happened? How did you feel afterwards?

I had an argument with my husband. We wanted to kill each other. I feel scared and unloved. I feel shitty about it

Submitted

Think about the circle of your friends. How do you feel when you meet them?

I feel horrible when I meet them. I feel that I'm worth nothing. I feel like a piece of trash

Submitted

Imagine you are in a room full of strangers. You were invited to the event but you know no one. How does it make you feel?

I'm scared. I feel awful when I need to be around people. I feel horrible to even think about this scenario

Submitted

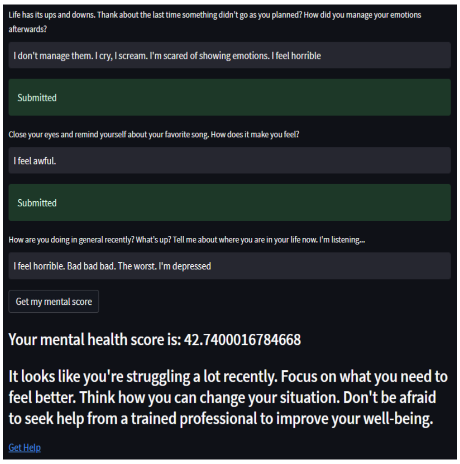
Recall the last moment when you achieved something. Did you have any recently? How did it make you feel?

I felt...

Figure 2: Example of user inputs in the En-for-Motion Chatbot.

Each answer has a predicted score assigned to it and a predicted emotion label. The score is normalized. The mean of all 8 scores is the final mental health score of the user. The range of possible scores derived from the training data used in the final model is 62.900986. The range is the difference between the highest predicted (most positive sentence in the training data) score and the lowest predicted score (most negative sentence in the training data).

Scores above 65 are ’above good’, scores between 65 and 50 are considered as ’good’, scores between 50 and 46 are ’bad’, scores between 46 and 41 - ’below bad’. Finally, the scores below 41 are considered ’critical’. To assign the positive ’good’ threshold, the chatbot was fed with all answers as ’good’ what gave the final score 63.713654. Answering to all questions as ’bad’ gives a score of 45.62471. If responding to each question with the word ’horrible’, the final score is 40.37927. These experiments constructed the threshold values for the chatbot’s response to the user’s mental health score (see Figure 3).



Life has its ups and downs. Thank about the last time something didn't go as you planned? How did you manage your emotions afterwards?

I don't manage them. I cry, I scream. I'm scared of showing emotions. I feel horrible

Submitted

Close your eyes and remind yourself about your favorite song. How does it make you feel?

I feel awful.

Submitted

How are you doing in general recently? What's up? Tell me about where you are in your life now. I'm listening...

I feel horrible. Bad bad bad. The worst. I'm depressed

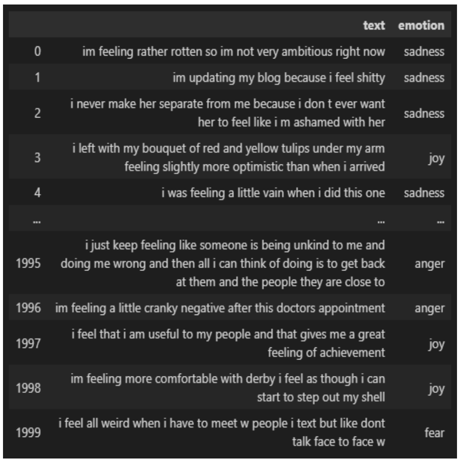
Your mental health score is: 42.7400016784668

It looks like you're struggling a lot recently. Focus on what you need to feel better. Think how you can change your situation. Don't be afraid to seek help from a trained professional to improve your well-being.

Figure 3: Example of the final mental health score calculated based on the user’s inputs.

2.1 Chatbot Data

En-for-Motion Chatbot and all the model experiments used Emotion Dataset for training. The dataset is open-source (https://github.com/dair-ai/emotion_dataset). It has already been preprocessed based on the approach described in the Saravia, et al. 2018 paper. Emotion Dataset is a dataset of English Twitter messages labeled for 6 basic emotions: anger, fear, joy, love, sadness, and surprise.



	text	emotion
0	im feeling rather rotten so im not very ambitious right now	sadness
1	im updating my blog because i feel shitty	sadness
2	i never make her separate from me because i don t ever want her to feel like i m ashamed with her	sadness
3	i left with my bouquet of red and yellow tulips under my arm feeling slightly more optimistic than when i arrived	joy
4	i was feeling a little vain when i did this one	sadness
...	...	...
1995	i just keep feeling like someone is being unkind to me and doing me wrong and then all i can think of doing is to get back at them and the people they are close to	anger
1996	im feeling a little cranky negative after this doctors appointment	anger
1997	i feel that i am useful to my people and that gives me a great feeling of achievement	joy
1998	im feeling more comfortable with derby i feel as though i can start to step out my shell	joy
1999	i feel all weird when i have to meet w people i text but like dont talk face to face w	fear

Figure 4: Examples of the Emotion Dataset used for training the chatbot model to assign emotion scores to the user’s answers.

There are 3 experiments conducted with Emotion Dataset. Data for Model1 has 18000 examples in total. It is split with 16000 and 2000 samples for the train and test set accordingly. Data for Model2 is extended and balanced by each class which is 10000 samples per class with the total data size of

Model	Test Accuracy	Val Accuracy	Epoch
Model1	94.95%	94.44%	9
Model2	**98.65%**	97.43%	5
Model3	98.55%	97.81%	3

Table 1: Performance of models used for training the En-for-Motion Chatbot. Model3 is the final model used for deploying Chatbot API.

60000. It was split 80/20. The train set has 48000 samples and test set includes 12000 samples.

The final dataset for Model3 (used in the final version for deployment of the chatbot) is balanced. It has 10000 samples per emotion label. It was split 80/20. The train set has 48000 samples and test set includes 12000 samples. This data setting is balanced because, to get the final score, the chatbot program puts 6 labels into 2 buckets: positive (joy and love) and negative (sadness, fear, anger and surprise).

2.2 Chatbot Model Architecture

The model used to train the En-for-Motion Chatbot has the following architecture. The Embedding layer is extracted from token based text embeddings trained on English Google News 7B corpus. The embeddings are accessed by TensorflowHub API (https://tfhub.dev/google/tf2-preview/nnlm-en-dim50-with-normalization/1). Text embeddings are based on feed-forward Neural-Net Language Models (Bengio, et al. 2003) with pre-built OOV. They map text to 50-dimensional embedding vectors. Next, model is fed with the Dense layer with ReLU activation function and L2 regularizer of 0.001 learning rate.

The training settings are the same in 3 experiments conducted to determine which dataset performs the best. The models use Adam optimizer with 0.001 learning rate. They are trained for 10 epochs, with batch size of 256 and early stopping. The models restore the best weights from each experiment. Each model uses validation split 0.2. The data is shuffled.

2.3 ER Chatbot Model Experiments

The Chatbot experiments investigate which data setting gives the best performance. Model2 turns out to have the best accuracy and the lowest loss on the test and validation set (see Table 1). Model2 is fed with the data where each emotion label has the same number of 10000 samples. However, as described above, the mental health score was calculated in the binary fashion - positive emotions vs negative emotions.

Therefore, Model2 is treated as imbalanced and Model3 as balanced. Surprisingly, Model2 is best performing with the accuracy of 98.65% on the test set. Model2 restores the best performing weights from epoch 5 with the accuracy of 97.43% on the validation set. Model3 performs with 98.55% accuracy on the test set. Weights of Model3 are restored from epoch 3 with the validation accuracy of 97.81% (see Table 1).

Model1 is treated as a baseline and it has much smaller sample size describe in the Chatbot Data subsection. Its performance is included in the Table 1.

Unfortunately, the RST experimental model did not bring expected results. Its performance was not taken into consideration. The assumption is that the data labeled for discourse relations was not of a high quality (see Figure 5). The results are not closely discussed and further research needs to be done to investigate if RST features can increase models' performance. The author believes that data curation and proper annotation of RST relations are necessary to use this feature and improve ER models' performance.

2.4 En-for-Motion Model Choice Motivation

Model3 is chosen to calculate the final mental health score in the Chatbot API, even though Model2 outperformed Model3. However, while calculating the score, Model3 gives a wider range between the most positive and the most negative score for the user's answer. Consequently, it is assumed that Model3 trained on the balanced data differentiates better between positive and negative responses. The understanding is that wider range of the answers' score allows for more fine-grained division between the levels of users' mental health and therefore, is more precise.

3 Discourse Relations and Emotions

3.1 RST and Emotion Data

To analyze the RST-Emotion correlations, the author uses Fairy Tales dataset (Alm, 2008). It includes 176 fairy tales manually annotated with 3 emotion labels (Joy, Anger,Fear, Sadness) and Neutral, 5 labels in total. However, the RST discourse parser is applied only to 95 fairy tales due to the model error. The final data includes 1125 sentences.

Label	Count	Total
NEUTRAL	3174	5725
no agreement	1426	
JOY	474	
SADNESS	322	
ANGER	203	
FEAR	126	

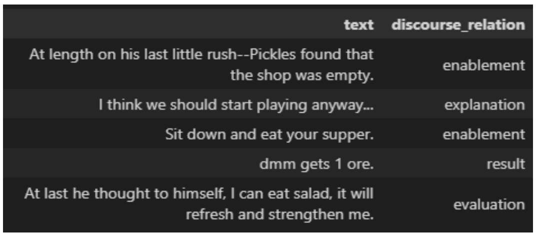
Table 2: Distribution of emotion labels (and neutral) in the final dataset. 'No agreement' labels are the cases when less than 75% of annotators agreed about the label.

Table 2 presents the distribution of emotion labels. After parsing, Fairy Tales dataset is labeled additionally for 16 unique RST discourse relations. The analysis though was focused on the top 3 and top 5 RST relations (see Table 3).

3.2 Discourse Relations in Dialogue Data

To answer the question if discourse relations can improve ER in NLP, the author attempts to fine-tune a BERT model for discourse relation classification. The STAC dataset is used for the model and is available open-source (https://www.irit.fr/STAC/corpus.html). It is a corpus of strategic chat conversations manually annotated with negotiation-related information, dialogue acts and discourse structures in the framework of Segmented Discourse Representation Theory (SDRT). For the purpose of this project, only discourse relation feature is used as a label. The dataset was developed within the context of the STAC (Strategic Conversation) project supported by the European Research Council (Asher, et al. 2016).

The STAC dataset is merged with Fairy Tales dataset and the discourse labels from both datasets are grouped to balance the data. The attempt to train the model does not bring expected results. The STAC dataset is useful only in the context of the conversation. Chunks of the dialogues shuffled with the Fairy Tales dataset do not convey the meaning of a discourse. Therefore, the author does not focus on this experiment in the paper.



text	discourse_relation
At length on his last little rush--Pickles found that the shop was empty.	enablement
I think we should start playing anyway...	explanation
Sit down and eat your supper.	enablement
dmm gets 1 ore.	result
At last he thought to himself, I can eat salad, it will refresh and strengthen me.	evaluation

Figure 5: Examples of the experimental discourse relation dataset (Fairy Tales and SCAT dataset merged).

RST Relation	Count
Contrast	2114
Elaboration	1936
Evaluation	1014
Enablement	211
Attribution	204

Table 3: Distribution of top 5 RST relations.

Nuclearity	Count
Satellite	4994
Nucleus	731

Table 4: Distribution of nuclearity.

3.3 ER Model for Fairy Tales

Zad and Finlayson (2020) paper uses the model developed by Kim et al. (2010) and improves its performance by 7.6 F1 points on average. The Zad and Finlayson's model corrected wrong and missing terms in relevant emotion lexicon, WordNetAffect http://corpustext.com/reference/affect_wordnet.html. The code and data for Zad and Finlayson (2020) is open-source and was reproduced to conduct the analysis of RST-Emotion correlations.

3.4 RST Model

The RST model used to assign discourse labels to Fairy Tales data is developed by Li et al. (2014) https://github.com/intfloat/TextLevelDiscourseParser. The model is based on MST Parser https://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html but improved with the dependency parsing techniques. The idea of the model is that discourse structure includes EDUs that are linked by the binary, asymetrical relations. The model assigns so called dependency relations to 2 EDUs: dependent and head. The links between the text spans are paired in the following manner: dependent-head. This aligns with the nucleus-satellite relation in RST theory.

3.5 RST Parser Experiments

The output of the RST Li parser seemes to not entirely capture the long documents' relations between the EDUs. The distribution of relations is very imbalanced. Different length of texts and different sizes of EDUs are applied to the RST parser to see if shorter sentences have better performance. The sentences in Fairly Tales test set are manually

Emotion Label	Nuclearity	Count
JOY	423	51
SADNESS	289	33
ANGER	174	28
FEAR	105	21

Table 5: Distribution of satellites and nuclei in each emotion label.

segmented into smaller chunks. Long fairy tales were divided into shorter paragraphs. The experiment shows that shorter chunks do not help the model performance. Model is confused and assigns the same discourse relations to most EDUs (Elaboration, Attribution, Evaluation and Contrast). It is assumed that the problem comes from the training data for the RST parser. Training data includes much smaller text spans than sentences in Fairy Tales dataset. Another reason is out-of-domain training data. The RST model was trained on short news articles, therefore, it does not predict well on long narrative fairy tales with dialogues.

RST Li parser results are compared to the RST Finder model results https://github.com/EducationalTestingService/rstfinder to validate Li parser. The author concludes that both models' results do not differ significantly. RST Finder model assigns mostly the same group of RST discourse relations as the Li model. Both models are not capable of assigning more fine-grained RST relations. RST Li parser is the final choice for the analysis.

The experiments with RST models leads to an interesting finding. Fairy Tales data is a mixture of long narrative sentences and dialogues with short question-answer turns. The model seems to better capture the correct RST relations in dialogue parts, even though the intuition is that story telling and developing the narration conveys more discourse structure meaning. It confirms that the narrative parts can benefit from RST parser trained on longer text chunks and dialogue parts will perform better when parsers are trained on conversational data.

3.6 Analysis - RST and Emotion Theory on Fairy Tales

The analysis of RST relations, nuclearity feature (nucleus or satellite) and emotion labels conducted on the Fairy Tales dataset has few major conclusions.

The most frequent emotion label used in the narrative texts is JOY followed by SADNESS, then ANGER and very few instances of FEAR. Table 5 shows the distribution of nuclearity in emotion labels. It might seem to follows the trend from the general distribution of emotion labels. However, the probability density function showed that SADNESS is more likely to occur as satellite than as nucleus. SADNESS is more likely to have a subordinate role in the discourse structure (satellite) than JOY. To ensure the validity of this assumption, the additional analysis is conducted. Sentences were divided into Nuclei and Sattelites sets. Positively polarized words (such as "great", "beautiful") occur more often in Nuclei than in Satellites. In contrast, the negatively polarized words (such as "cry", "poor") have are more often in Satellites.

Contrast was assigned to highest number of sentences, followed by Elaboration, Evaluation, Attribution and Enablement. Distribution of top 5 relations is normalized using probability density function. Attribution has a very similar probability for 2 most frequent emotion labels: SADNESS and JOY (around 0.33-0.35). However, it was not assigned to any sentence from the FEAR category. Enablement has a very high probability to occur in the ANGER sentences (0.3) comparing to JOY (0.39). Though, the same relation Enablement has much lower probability to occur with sentences labeled for SADNESS.

Evaluation relation in ANGER and SADNESS sentences has similar probability score (around 0.24 and 0.27 respectively). The normalized graphs show also that even though Contrast is a top relation and JOY is a top emotion label, Elaboration is significantly more often used with the JOY label than with any other emotion. Contrast has the highest probability score in FEAR. The least occurring emotion label has the highest probability score of the most frequently occurring RST relations in the whole dataset.

The analysis looked as well into the 5 least frequently used RST relations: Joint, Explanation, Comparison, Cause and Textual. 3 of them occur in SADNESS sentences, 2 in JOY and 1 in ANGER. SADNESS category has the most diversity in the RST relations. SADNESS sentences are labeled with 13 out of 17 unique RST categories, where 41.3% belong to Contrast (comparing to 41.1% on JOY). Each emotion is analyzed also in terms of their nuclearity distribution per RST relation. FEAR and ANGER sentences, the least frequently

Emotion Label	Contrast and Nucleus
JOY	22
SADNESS	9
ANGER	16
FEAR	12

Table 6: Distribution of satellites and nuclei in each emotion label.

annotated emotions, are more likely to have Nuclei features than SADNESS or JOY.

3.7 RST Parsing Error Analysis

Multiple discourse parsing models vary in terms of the configurations and options, as well as the input format. The RST Li model (2014) is chosen for the analysis, however, the other models are checked as potential alternatives, the HILDA discourse parsing model https://github.com/NLPbox/hilda-docker and the MEGA DT discourse parser https://github.com/nlpat/MEGA-DT.

Discourse parsing remains a problematic issue in NLP. Many models are not publicly available and are not pre-trained. The data needed for training is not rarely unavailable open-source. Outputs of RST parser differ in form. Some models offer the RST tree builders but specified for a certain model. They cannot be used universally. The process of examining the agreement between the outputs is complicated and time-consuming. Additionally, the models often have EDU segmenters specific for the model or they are not provided at all. Evaluation of the impact of discourse relations on NLP tasks remains challenging and complex.

4 Conclusions

The En-for-Motion Chatbot presents a promising avenue for the assessment of mental health. By leveraging the power of AI, this chatbot can offer personalized and mental health support to individuals worldwide. With the increasing demand for mental health services, chatbots like En-for-Motion can offer a scalable and cost-effective solution to mental health diagnosis and therapy. While there is still much work to be done, this study highlights the potential of chatbots in mental health, paving the way for further research in this exciting field. Ultimately, the successful integration of AI and mental health has the potential to revolutionize mental health care, making it more accessible and efficient for those who need it most.

The deep analysis of dependencies between emotion labels and the discourse relations provides valuable insights into the ways in which emotions are expressed and conveyed in language. Findings presented in the paper can help in better understanding of the underlying mechanisms that shape the emotional responses to language. Experiments with numerous RST parser and data setting confirmed the need for more thorough research with in-domain focus for discourse relations, as well as more open community approach with pre-trained models and open-source data. Ultimately, this research contributes to a deeper understanding of the ways in which language reflects and shapes our emotional experiences, and underscores the need for continued investigation into the complex interplay between emotion and language.

References

Alm. 2008. Affect in Text and Speech. lrc.cornell.edu.

Alm, Cecilia Sproat, Richard. (2005). Emotional Sequencing and Development in Fairy Tales. 668-674. 10.1007/1157354886.

Asher, N., Hunter, J., Morey, M., Benamara, F. S. Afantenos (2016). Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In The Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, pp. 2721-2727, Portorož.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3:1137-1155.

Bosselut, A., Celikyilmaz, A., He, X., Gao, J., Huang, P., Choi, Y. (2018). Discourse-Aware Neural Rewards for Coherent Text Generation. ArXiv, abs/1805.03766.

Busso, C., Bulut, M., Lee, CC. et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources Evaluation 42, 335 (2008). https://doi.org/10.1007/s10579-008-9076-6

Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. NAACL.

Edwards, D. (1999). Emotion discourse. Culture Psychology, 5(3), 271–291. https://doi.org/10.1177/1354067X9953001

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169–200.

Fahrni, A., Strube, M. (2014). A Latent Variable Model for Discourse-aware Concept and Entity Disambiguation. EACL.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated Essay Scoring with Discourse-Aware Neural Models. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 484–493, Florence, Italy. Association for Computational Linguistics.

Ghazvininejad, M., Karpukhin, V., Celikyilmaz, A. (2021). Discourse-Aware Prompt Design for Text Generation. ArXiv, abs/2112.05717.

Ghosh, S., Srivastava, H., Umesh, S. (2022). A Discourse Aware Sequence Learning Approach for Emotion Recognition in Conversations. ArXiv, abs/2203.16799.

Hans Kamp. 1981. A theory of Truth and Semantic Representation in J. Groenendijk, TH. Janssen and M. Stokhof(eds.) Formal Methods in the Study of Language, Part I. Mathematisch Centrum, Amsterdam. pages 277-322

Heilman, M., Sagae, K. (2015). Fast Rhetorical Structure Theory Discourse Parsing. ArXiv, abs/1505.02425.

Hernault, Hugo Prendinger, Helmut duVerle, David Ishizuka, Mitsuru. (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. Dialogue Discourse; Vol 1, No 3 (2010). 1. 10.5087/dad.2010.003.

Huang, Y., Fang, M., Cao, Y., Wang, L., Liang, X. (2021). DAGN: Discourse-Aware Graph Network for Logical Reasoning. NAACL.

Huber, P., Carenini, G. (2020). MEGA RST Discourse Treebanks with Structure and Nuclearity from Scalable Distant Sentiment Supervision. ArXiv, abs/2011.03017.

Lech, M., Stolar, M.N., Best, C., Bolia, R.S. (2020). Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. Frontiers in Computer Science.

Landowska, Agnieszka (2019). Uncertainty in emotion recognition. Journal of Information, Communication and Ethics in Society 17 (3):273-291.

Lee, K., Han, S., Myaeng, S. (2018). A discourse-aware neural network-based text model for document-level text classification. Journal of Information Science, 44, 715 - 735.

Li, S., Wang, L., Cao, Z., Li, W. (2014, June). Text-level discourse dependency parsing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 25-35).

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. IJCNLP.

Liang-Chih Yu, Chung-Hsien Wu, Fong-Lin Jang, Psychiatric document retrieval using a discourse-aware model, Artificial Intelligence, Volume 173, Issues 7–8,2009,Pages 817-829,ISSN 0004-3702,https://doi.org/10.1016/j.artint.2008.12.004.

Mann, W.C., Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text Talk, 8, 243 - 281.

Märkle-Huss, J., Feuerriegel, S., Prendinger, H. (2017). Improving Sentiment Analysis with Document-Level Semantic Relationships from Rhetoric Discourse Structures. HICSS.

Meade, G.G. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.

Mihaylov, T., Frank, A. (2019). Discourse-Aware Semantic Self-Attention for Narrative Reading Comprehension. ArXiv, abs/1908.10721.

Miltsakaki, Eleni Prasad, Rashmi Joshi, Aravind Webber, Bonnie. (2004). The Penn Discourse Treebank.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better Document-level Sentiment Analysis from RST Discourse Parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Plutchik, R. 1980 “Measurement implications of a psychoevolutionary theory of emotions”, in: K.R. Blankstein ; P. Pliner ; J. Polivy (eds.). Assessment and modification of emotional behavior. New York, Plenum.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ACL.

Purpura, A., Masiero, C., Silvello, G., Susto, G. (2019). Feature Selection for Emotion Classification. IIR.

Multilingual Dependency Parsing with a Two-Stage Discriminative Parser R. McDonald, K. Lerman, and F. Pereira Tenth Conference on Computational Natural Language Learning (CoNLL-X) (2006)

Reinhard Pekrun. (2021) Teachers need more than knowledge: Why motivation, emotion, and self-regulation are indispensable. Educational Psychologist 56:4, pages 312-322.

Saravia Elvis, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Strapparava, C and Valitutti A. (2004). WordNet-Affect: an affective extension of WordNet. Proceedings of the 4th International Conference on Language Resources and Evaluation 1083–1086.

Xiong, H., He, Z., Wu, H., Wang, H. (2019). Modeling Coherence for Discourse Neural Machine Translation. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 7338-7345.

Xu, J., Gan, Z., Cheng, Y., Liu, J. (2020). Discourse-Aware Neural Extractive Text Summarization. ACL.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2949–2958, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Z. Shi and M. Huang, “A deep sequential model for discourse parsing on multi-party dialogues,” in AAAI 2019, vol. 33, no. 01, pp. 7007–7014.

Zad, Samira; Finlayson, Mark A., 2020, "Data and Code for "Systematic Evaluation of a Framework for Unsupervised Emotion Recognition for Narrative Text"", https://doi.org/10.34703/gzx1-9v95/03RERQ, FIU Research Data Portal, V1

A Discourse Relations Background

A.1 Discourse in Computation

Discourse in itself has always been in the foundation of computational theories of language. DRT (Discourse Representation Theory) was the first attempt to formalize the semantic model of text as a discourse (Kamp, 1981). It aimed to be applied for technical purposes and had its applications in discourse understanding by systems. Now discourse theories are widely used in NLP and discourse-aware models tend to achieve better performance in many NLP tasks. Discourse structure is deployed in the fields such as machine translation (Xiong et al., 2019), text summarization (Xu et al., 2020; Cohan et al., 2018), text generation (Bosselut et al., 2018; Ghazvininejad et al., 2021), text classification (Lee et al., 2017), document retrieval (Yu et al., 2009), automated essay scoring (Nadeem et al., 2019), logical reasoning question answering (Huang et al. 2021), narrative reading comprehension (Mihaylov and Frank, 2019) or enitity disambiguation in texts (Fahrni and Strube, 2014). The feature of discourse structure that the paper focuses on is extracting discourse relations between spans of text. There are 2 main corpora for parsing a discourse into semantically related relations of sequences of text: RST Disourse TreeBank (Carlson et al., 2001) and Penn Discourse Tree Bank (PDTB) (Miltsakaki et al., 2004). For the purpose of the presented analysis, RST Discourse Treebank was used in this paper.

A.2 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) (Mann and Thomson, 1988) serves as an explanation of which elements of a text make it coherent and how these elements are connected to each other. RST implies that each part of a coherent text has a specific function and a reason for its presence that is evident to receivers. RST represents a text as a tree that has hierarchical structure. Each leaf of a tree corresponds to a text span called Elementary Discourse Unit (EDU). The theory derives from the notion that two EDUs detected in a text are relative to each other and each has a specific role in a coherent discourse. The texts spans possess Nucleus-Satellite relations which points the direction in the hierarchy between two text spans. Nucleus is considered as a claim, a reason for the text spans to exists in a coherent discourse. Satellite will play a role of a subordinate part giving evidence for confirming that the claim in the discourse structure is valid and makes the text coherent. The order of this hierarchical relations is not fixed but there are patterns of order that follow certain discourse relations. RST accepts multinuclear relations when one text span in a pair does not hold a more important role than the other one. These relations were not taken into consideration for the purpose of this analysis. In the paper, 17 unique discourse relations were analyzed and compared to their nuclearity feature (Nucleus vs Satellite) and emotion labels annotated per sentence in the fairy tale dataset. The available RST parsing models comes with different configurations and are usually divided into two parts: segmenters and parsers. Segmenters are responsible for breaking down a text to EDUs, whereas parsers assign the correct discourse relation to a specific pair of EDUs within a nuclearity relation between them. This raises a question about the appropriate way to segment a text into EDUs. To conduct the analysis of this paper, so called text-level parsing model was utilized. Parsing process can also be carried out on the sentence level connecting the clauses within a sentence by assign the discourse relations between them. Since the fairy tale dataset was annotated as an emotion label per sentence, each sentence was treated as an EDU in the discourse. This way of conducting the parsing process is called the inter-sentential discourse parsing.

A.3 Discourse in Emotion Recognition (ER)

Even though discourse feature started being applied to Emotion Recognition models in the recent years, most of work that has already been done focuses on conversational data and ER for dialogues. Narrative texts or domain-specific texts, such as online posts, have not been explored yet with the use of discourse-aware features for ER. However, there has been a significant interest in Emotion Recognition in Conversation (ERC) and Emotion Recognition in Multi-Party Conversation (ERMC). These models tend to find correlations between conversational and sequential context in conversational texts. The premise of these models is that there is a dependency between how the turns in a dialogue taken by its participants and the relations governed by the discourse structure. Conversational data for ERC and ERMC consists of speaker-specific features with annotations for the interlocutor taking a turn in the dialogue. Some of the corpora used for ERC and ERMC models are: IEMOCAP (Busso et al., 2008), MELD (anger, disgust, sadness, joy, neutral, surprise, fear) (Poria et al., 2019) or DailyDialog (Li et al., 2017). These annotated corpora usually rely on segmenting an EDU as a sentence which aligns with the approach taken in the paper. One of the reasons is the length of an average sentence in a dialogue. Discourse features, such as relations between EDUs, were also adopted to the sentiment analysis models improving the performance by parsing documents with the use of RST.

A.4 Emotion Theories for Computation

Many emotion theories have been established and developed before natural language was used for detecting emotions. Ekman's theory of emotions proposed in the paper (Ekman, 1992) suggests that each emotion has distinctive features: signal, physiology and antecedent events. Thanks to that, emotions can be categorized into the set of basic primitives and built upon expressing more affective phenomena. Ekman's theory of emotions is the most frequently used set of emotions in computer science (Purpura et al., 2019). The theory establishes 7 emotions in total: anger, fear, disgust, joy, sadness, surprise (+/-). For the purpose of this analysis, the set of emotions was limited to four different labels: JOY, SADNESS, ANGER and FEAR, following the approach taken in Zad and Finlayson (2020). Plutchik's theory of emotions represents a theory that combines categorical and dimensional approach to emotion theories (Plutchnik, 1980). Plutchnik's wheel of emotions presents the idea of dimensionality of emotions by categorizing them in different intensity groups while keeping the emotions grouped into categorical parts. Future Work section discusses a possible direction for aligning dimensions of emotions with dimensions of discourse relations. The discourse-aware models for ER mostly focus on applying discourse dependency parsers and strictly improving the performance of models providing only the quantitative results of the experiments. The work is not focused on the thorough analysis leading to understanding the dependencies between emotion labelling and discourse relations. Markle-Huß et al. (2017) presents the conclusion of the correlation between nuclearity and sentiment classification when predicting the stock market reaction. Based on the conducted experiments they prove that N (nucleus) and NS (nucleus-satellite) relations of the RST tree bear the most significant information for sentiment classification. However, their work is one of few that provides this sort of qualitative analysis. Moreover, their work focuses on predicting stock returns subsequent to financial disclosures and it's heavily related to this domain. Similarly to the approach in Markle-Huß et al. (2017), the analysis in this paper aims to propose weighting the emotion labels from the RST tree when applying RST parsing to ER models.