Rationalization in Natural Language Processing

A Survey and Introduction to Rational AI

- Purvil Patel
- CMPE-255 Data Mining

Abstract

- **Improvement in NLP and Trade-off:** Recent advancements in deep learning have significantly enhanced the performance of various NLP tasks such as translation, question-answering, and text classification. However, this improvement has come at the cost of model explainability.
- **Challenge of Black-box Models:** These black-box models are challenging to understand, especially regarding their internal workings and the process they use to reach a conclusion.
- **Rationalization as a Solution:** To address this, rationalization has emerged as a more accessible and intuitive technique for explainability in NLP. It provides natural language explanations (rationales) for a model's outputs, making them understandable even to non-technical users

Introduction

- **NLP Growth and Applications:** Over the past decade, NLP has seen tremendous growth and has been applied in various fields such as text classification, fact-checking, machine translation, and text-to-speech.
- **Explainability Challenge:** Despite its widespread application, a significant challenge NLP faces is explainability. The shift from traditional white-box techniques to black-box models in deep learning has enhanced performance but reduced transparency.
- **Impact and Risks of Lack of Explainability:** This lack of transparency can erode trust between humans and AI systems, posing risks in critical areas such as healthcare, finance, and law. For instance, a medical recommendation system without transparent decision-making processes can potentially lead to harmful outcomes.
- **Research Focus on Interpretability and Explainability:** There has been a focused effort in research to make models more interpretable and complete, with explainability being a crucial aspect. The goal is to enable end-users to understand how a model arrives at its results

Background

- **Definition:** Rationalization involves providing natural language explanations to justify a model's prediction. These explanations, known as rationales, highlight the input features influencing the model’s decision.
- **Human-Comprehensible Technique:** It is an approachable technique for non-technical users, as it allows them to understand the model's decision-making process simply by reading the rationale.
- **Difference from Other Techniques:** Unlike other explainability methods, rationalization focuses on local explanations, which are specific to individual predictions, making it more detailed and specific

Key Contributions

- **Comprehensive Analysis:** This research represents the first extensive survey of rationalization literature in NLP, covering works from 2007-2022.
- **Rational AI (RAI):** Introduction of a new subfield in Explainable AI, termed Rational AI, aimed at advancing rationalization techniques.
- **Survey Contents:** The survey includes methods, evaluations, code, and datasets used in various NLP tasks employing rationalization.
- **Insights and Future Directions:** The paper discusses insights, challenges, and future directions in the field, pointing to promising research opportunities

Definitions

- **Black-box Model:** A type of AI model whose internal workings are not visible or understandable to the user.
- **Interpretability:** The ability to understand the internals of a model.
- **Explainability:** The capacity of a model to provide understandable explanations of its operations and decisions.
- **Rationalization:** Providing natural language explanations for model predictions.
- **NLP Assurance:** Ensuring the reliability and trustworthiness of NLP systems.

Methodology

- **Emphasis on Rationalization:** The survey selectively focuses on publications that contribute specifically to rationalization within the field of NLP.
- **Time Span Coverage:** The survey includes studies and works from 2007 to 2022, offering a comprehensive historical perspective of the advancements in rationalization.
- **Domain Inclusivity:** Publications from various domains within NLP are considered to ensure a diverse and thorough understanding of rationalization applications across the field.

Rationalization Techniques

- **Categorization by Task:** An outline of different rationalization techniques, organized according to the NLP tasks they are applied to.
- **Technique Descriptions:** Brief explanations of each technique, highlighting their unique contributions to enhancing explainability in their respective NLP tasks.

Extractive Rationalization

- Extractive Rationalization: Involves extracting important features or sentences from the input data as rationales to support the model's prediction.
- Usage in NLP: Commonly used to identify and highlight key parts of the input data that significantly influence the model's decision-making process.
- Example Models: Techniques often involve traditional machine learning models or simpler neural network architectures.

Abstractive Rationalization

- Abstractive Rationalization: A generative task where novel sentences are created using new words or paraphrasing existing sentences to explain a model's predictions.
- Generative Models: Typically involves advanced language models like T5 or GPT for generating these novel explanations.
- Advantages: Provides more flexible and comprehensive explanations by generating entirely new content rather than being limited to existing input data

Machine Reading Comprehension - An Overview

- **Definition:** AI's ability to read and interpret text.
- Importance in today's digital landscape.
- **Applications:** Chatbots, automated customer support, and more.

Technological Milestones in MRC

- Evolution from simple algorithms to complex neural networks.
- Improvements in understanding context and ambiguities.

Key Paper in MRC

Table 3. Machine Reading Comprehension Papers

Paper	Name	Year	Explanation	Models	XAI Metric	Dataset	Code
(Sharp et al. 2017)	-	2017	Extractive	TF-IDF, FFNN	-	AI2 Science, Aristo Mini	-
(Ling et al. 2017)	-	2017	Extractive	LSTM, Seq2Seq	-	AQuA	✓
(Mihaylov et al. 2018)	OpenBookQA	2018	Abstractive	BiLSTM Max-out	-	OpenBookQA,	✓
(Xie et al. 2020)	WorldTree V2	2018	Abstractive	TF-IDF, BERT	-	WorldTree V2	✓
(Lakhotia et al. 2021)	FiD-Ex	2021	Extractive	T5, BERT-to-BERT	-	Natural Questions	-

Challenges and Future Directions

- **Current Limitations:** Addressing challenges such as understanding context and ambiguity in text.
- **Future Prospects:** Exploring potential advancements and the integration of more sophisticated AI techniques.

Commonsense Reasoning

- **Rationalization Techniques:** Description of techniques that enable models to provide natural language explanations for predictions involving everyday knowledge and commonsense reasoning.
- **Key Resources:** A list of pivotal papers, models, and datasets in the domain of commonsense reasoning, highlighting how rationalization improves the comprehensibility and reliability of the outcomes.

Breakthroughs in Commonsense Reasoning

- Overview of major contributions.
- The role of commonsense reasoning in complex inference.

Key Papers

Table 4. Commonsense Reasoning Papers

Paper	Name	Year	Explanation	Models	XAI Metric	Dataset	Code
(Ehsan et al. 2018)	-	2018	Extractive	LSTM, Seq2Seq	-	-	-
(Rajani et al. 2019)	CAGE	2019	Abstractive	GPT, BERT	-	CoS-E, CommonsenseQA	✓
(Majumder et al. 2021)	RExC	2021	Extractive	Transformer	-	ComVE, e-SNLI, COSe, e-SNLI-VE, VCR	-
(Tang et al. 2021)	DMVCR	2021	Extractive	LSTM, BERT	-	VCR	✓

Overcoming Challenges in Common Reasoning

- Dataset Limitations: Discussing the hurdles in creating datasets that effectively capture commonsense knowledge.
- Prospective Improvements: Potential future developments to enrich commonsense reasoning capabilities in AI models

Natural Language Inference

- **Explainability in Inference:** Exploring techniques that explain how models determine relationships between sentences or phrases in the context of natural language inference.
- **Important Contributions:** Identification of significant research, models, and datasets that have shaped the application of rationalization in natural language inference.

Evolving Research in NLI

- Review of significant studies and datasets.
- Challenges in creating reliable NLI models.

Key Papers

Table 5. Natural Language Inference Papers

Paper	Name	Year	Explanation	Models	XAI Metric	Dataset	Code
(Camburu et al. 2018)	e-SNLI	2018	Abstractive	BiLSTM, Seq2Seq	-	e-SNLI	✓
(Kumar and Talukdar 2020)	NILE	2020	Abstractive	GPT-2, RoBERTa	-	e-SNLI	✓
(Wiegreffe et al. 2021)	-	2020	Abstractive	T5	-	CoS-E, SNLI	✓

Future of NLI

Challenges in Explainability: Addressing the need for faithful and interpretable explanations in NLI models.

Emerging Techniques: Exploring new frameworks like NILE for generating natural language explanations in NLI tasks

Fact-Checking

- **Techniques for Verification:** Detailing rationalization techniques that assist in explaining the process of verifying factual accuracy in text.
- **Crucial Developments:** Listing key research papers, models, and datasets that have been influential in applying rationalization to fact-checking tasks.

Advances in NLP Fact-Checking

- Overview of novel datasets and techniques.
- Addressing the challenges in claim verification.

Key Papers

Table 6. Fact-Checking Papers

Paper	Name	Year	Explanation	Models	XAI Metric	Dataset	Code
(Alhindi et al. 2018)	LIAR-PLUS	2018	Extractive	SVM, BiLSTM	-	LIAR-PLUS	✓
(Hanselowski et al. 2019)	-	2019	Extractive	BERT	-	FEVER	✓
(Atanasova et al. 2020)	-	2020	Extractive	DistilBERT	-	LIAR-PLUS	-
(Rana et al. 2022)	RERRFACT	2022	Extractive	RoBERTa, BioBERT	-	SCIFACT	-

Fact-Checking Challenges

- Complexities: Examining the difficulties in classifying claims correctly, especially with heterogeneous data sources.
- Advancing Fact-Checking: Speculating future improvements and the role of AI in enhancing the accuracy and reliability of fact-checking systems

Sentiment Analysis

- **Understanding Sentiment through Rationalization:** Description of how rationalization techniques elucidate the process of determining sentiment in text.
- **Influential Works:** A compilation of important papers, models, and datasets that have contributed to rationalization in sentiment analysis.

Rationalization Techniques in Sentiment Analysis

- Exploration of models for transparent predictions.
- Importance of justifiable outcomes.

Key Papers

Table 7. Sentiment Analysis Papers

Paper	Name	Year	Explanation	Models	XAI Metric	Dataset	Code
(Lei et al. 2016)	-	2016	Extractive	LSTM, RCNN	-	BeerAdvocate, AskUbuntu	✓
(Du et al. 2019)	CREX	2019	Extractive	CNN, LSTM	-	BeerAdvocate, MovieReview	-
(Strout et al. 2019)	-	2019	Extractive	RA-CNN, AT-CNN	-	MovieReview	-
(Yu et al. 2021)	A2R	2021	Extractive	BiGRU	-	BeerAdvocate, MovieReview	✓
(Antognini and Faltings 2021)	ConRAT	2021	Extractive	CNN, BiGRU	-	AmazonReviews, BeerAdvocate	-

Future Trends in Sentiment Analysis

- Challenges in Explainability: Addressing the need for models that provide explanations aligning with established domain knowledge.
- Prospective Developments: Discussing potential future advancements in making sentiment analysis models more credible and interpretable

Text Classification

- Definition: Explain NMT as the use of deep neural networks to translate text from one language to another.
- Importance: Discuss the role of NMT in breaking language barriers and enabling global communication.
- Evolution: Briefly trace the progression from traditional translation methods to neural network-based approaches.

Literature Review in NMT

- Paper Title: "Sequence to Sequence Learning with Neural Networks" by Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.
- Basic Info: Discuss the pioneering work of Sutskever et al. in applying sequence-to-sequence learning for NMT. This paper laid the groundwork for many modern NMT systems.
- Contributions: Explain how this paper contributed significantly to the field by introducing novel methods that improved the quality and efficiency of machine translation.
- Impact: Highlight the lasting impact of this paper on subsequent research and the development of more advanced NMT systems.

Technological Advancements in NMT

- Key Developments: Highlight significant technological advancements in NMT, such as the use of LSTM networks, attention mechanisms, and transformer models.
- Challenges: Discuss ongoing challenges in NMT like handling contextual nuances, idiomatic expressions, and maintaining translation accuracy across diverse language pairs.
- Future Prospects: Speculate on potential future developments in NMT, such as improved handling of low-resource languages and integration of cultural context.

Multiple Domains

- Versatility of Rationalization: Discussing how rationalization techniques are applied across multiple NLP domains, demonstrating their versatility and broad applicability.
- Cross-Domain Insights: Presenting a synthesis of key research findings, models, and datasets that showcase the use of rationalization in various NLP tasks, illustrating its adaptability and effectiveness in diverse contexts.

Conclusion

- **Overview of Findings:** A summary of the main insights and advancements in the field of rationalization in NLP as highlighted throughout the presentation.
- **Addressing Challenges:** Discussing the challenges encountered in rationalization and the importance of ongoing research in this area.
- **Future Directions:** Highlighting potential future directions and the evolving role of Rational AI in enhancing the explainability and trustworthiness of NLP models.