Advancing Identification of Causal Relations in Natural Language

by Pedram Hosseini

B.S. in Computer Engineering, September 2012, University of Guilan
M.S. in Computer Engineering, May 2016, University of Guilan

A Dissertation submitted to

The Faculty of
The School of Engineering and Applied Science
of the George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2022

Dissertation directed by

Mona T. Diab
Professor of Computer Science

The School of Engineering and Applied Science of The George Washington University certifies that Pedram Hosseini has passed the Final Examination for the degree of Doctor of Philosophy as of May 18, 2022. This is the final and approved form of the dissertation.

Advancing Identification of Causal Relations in Natural Language

Pedram Hosseini

Dissertation Research Committee:

Mona T. Diab, Professor of Computer Science, Dissertation Director

David A. Broniatowski, Associate Professor of Engineering Management & Systems Engineering, Committee Member

Abdou Youssef, Professor of Computer Science, Committee Member

Robert Pless, Patrick & Donna Martin Professor of Computer Science, Committee Member

Aylin Caliskan, Assistant Professor in the Information School, University of Washington, Committee Member

© Copyright 2022 by Pedram Hosseini
All rights reserved

Dedication

To my parents and my brother, Poorya, for their unconditional love and support.

Acknowledgments

Although words may not do my message justice, I would like to take this opportunity to express my gratitude to the people whose support was integral to completing this thesis and greatly enriched my doctoral studies at The George Washington University. Finishing my Ph.D. was not possible without having the support of the best advisors I could ask for, **Dr. Mona T. Diab** and **Dr. David A. Broniatowski**. I will forever be grateful to Mona for placing her trust in me and giving me an opportunity to pursue a doctorate degree in such a great institution and lab environment. Mona was always patient with my questions and passed on to me tools and principles that will always guide me in my scientific endeavors. She additionally taught me how important it is to make an impact on our community beyond pure scientific contributions. It was quite fortunate for me to get to know David early on during my doctoral studies and have a chance to do research that shows the great impact of Natural Language Processing (NLP) beyond immediate computer science applications to address broader societal issues. David showed me how one can stay humble despite great scientific achievements, and how to embrace new and different ideas with an open mind when these ideas are outside the 'familiar' territories. Like Mona, he always supported me when I needed any help and resources to achieve my research goals. I am grateful to Mona and David for their support, for giving me the freedom to explore, and for always being there with their knowledge, words of wisdom, and guidance. To Mona and David, I say thank you. I have always been proud to call myself your student.

I would also like to thank my committee members, **Dr. Abdou Youssef**, **Dr. Robert Pless**, and **Dr. Aylin Caliskan**, who were critical in shaping the contents and direction of my dissertation and made themselves available whenever I needed their guidance. They always provided me with constructive feedback, taught me how to remain focused on asking the correct questions in my research, and improve my thinking about communicating my message to make meaningful contributions to my research community. To all of you, I say thank you; it was a privilege to have you as my committee members on this journey. The

lessons I learned from you will be with me for the journeys ahead.

I will be forever grateful to my family for their unconditional love and support. I would like to start by thanking my parents without whom I would not have been the person I am today. I lost my father before he knew I started my Ph.D. studies. I spent many days at his office, in the airport of my hometown, Ramsar, when he patiently taught me how to write commands on a disk operating system for the first time back in the '90s—the time I started to grow an interest in computers. My father always had a funny story to tell with a beautiful smile. My brother and I made a promise to keep that smile on his face. I am sure seeing me finishing my Ph.D. in computer science makes him proud and happy. My mother always showed me how to be strong in times of difficulty, cheered me up and believed in me, and wholeheartedly supported me throughout my life. Leaving her to come to the United States to pursue my dreams was not easy after losing my father, but she never made me doubt even for a moment that I should follow my heart and pursue what I am passionate about. To my parents, I say, I love you and you are truly my rockstars.

I would also like to thank my one and only brother, Poorya—Dr. Hosseini, the first—for all he did for me. Poorya has been more than a brother to me. He gifted me with his experience, patience, mentorship, and unconditional support whenever I needed him and provided me with the invaluable lessons he learned himself the hard way. Poorya always reminded me of my ability to accomplish anything I put my mind to, pushed me to be a better version of myself, and taught me how to have a sight on the longer-term vision and the bigger picture. To him, I say, I love you and I hope I make you proud.

Last but not least, I would like to thank all my friends, mentors, lab mates, and collaborators whom I had the opportunity to get to know, make memories and work with, and learn from. I was fortunate to be a member of the CARE4LANG and Decision Making and Systems Architecture (DMSA) labs at The George Washington University. I learned many lessons through the fruitful discussions we had in all the lab meetings and journal club sessions. To all of you I say thank you, and the memories we made and the lessons I learned

will not be forgotten. I wish the best for each of you on your life journeys and look forward to staying in touch for years to come.

Abstract

Advancing Identification of Causal Relations in Natural Language

As the amount of available digital text has been growing exponentially, the development of tools and algorithms for processing this often-unstructured natural language is critical to providing scalable solutions for various applications. These algorithms are often classified under the broad field of Natural Language Processing (NLP). One major component of NLP is the identification of the semantic relationships within various constituents. The tools that can be used to identify causal relations are the ones that are relatively less advanced, despite their promise and potential to unlock great value across various applications.

In this thesis, we investigate three different applications where the identification of causal relations can solve major problems and, in turn, provide insight into where and how causal relation identification models can be improved to increase their performance in these specific application areas and beyond. By introducing new frameworks and building novel tools, we first demonstrate the utility of causal relations in the following areas, (1) methods for discerning between satire and false news, (2) analyzing and predicting the spread of various types of information online, and (3) developing computational tools for characterizing the capability of a given text in creating a coherent mental representation for readers. Following these demonstrations, we initially highlight the need to build more robust models for accurate identification of causal relations in natural language, following which we present a framework for building new commonsense knowledge-augmented models to identify causal relations and explanations in the text.

In summary, in this thesis, we initially demonstrate how enhanced causal relation models and tools can enable the automation of advanced tasks within three impactful application areas. These demonstrations, in turn, provide insight into how novel frameworks and tools can be developed to increase the power and efficacy of causal relation identification algorithms in these applications and beyond. Throughout the thesis and in each chapter, we

also comment on how our research can be used to guide the field through future developments of causal relation classification models and algorithms.

Table of Contents

Dedication . . . . . **iv**

Acknowledgments . . . . . **v**

Abstract . . . . . **viii**

List of Figures . . . . . **xiii**

List of Tables . . . . . **xv**

Abbreviations **1**

1 Introduction **3**

1.1 Motivation 3

1.2 Thesis outline 4

2 Identifying Nuances Between Satire and Fake News **7**

2.1 Introduction 7

2.2 Related work 9

2.3 Method 9

2.3.1 Semantic representation with BERT 10

2.3.2 Linguistic analysis with Coh-Metrix 10

2.4 Results and discussion 11

2.4.1 Classification between fake news and satire 11

2.4.2 Insights on linguistic nuances 13

2.5 Conclusion and future work 14

3 Does Causal Coherence Predict Online Spread of Social Media? **16**

3.1 Introduction 16

3.2 Related work 18
3.3 Method and analysis 21
3.3.1 Dataset and preprocessing 21
3.3.2 Measuring causal coherence 21
3.3.3 Analysis 22
3.4 Results and discussion 22
3.5 Conclusion and future work 25

4 Measuring Gist Inference Score in Text 28

4.1 Methods 30
4.1.1 Local vs. global indices 30
4.1.2 GisPy indices implementation 31
4.1.3 Computing Gist Inference Score (GIS) 34
4.2 Experiments 34
4.2.1 Benchmarks 35
4.3 Results and discussion 36
4.3.1 Testing robustness 38
4.4 Conclusion and future work 43

5 A Framework for Cause-Effect Relation Classification 44

5.1 Introduction 44
5.2 Definition of causal relations 46
5.2.1 Causal relation categorization 47
5.2.2 Causal relation identification tasks 48
5.3 History of automatic causal relation identification in text 49
5.3.1 Before 2000: hand-coding and rule-based methods 49
5.3.2 2000-2018: machine learning starts to play a role 52
5.3.3 2018-Present: the dominance of deep neural networks 61

5.4 Method 64
5.4.1 ATOMIC$^{20}_{20}$ to text 65
5.4.2 GLUCOSE to text 67
5.4.3 Continual pretraining 68
5.5 Benchmarks 69
5.5.1 CREST: A causal relation schema for text 71
5.6 Experiments 73
5.7 Results and discussion 76
5.7.1 Error analysis 80
5.8 Conclusion 82
Bibliography **84**

List of Figures

3.1 Normal probability of logarithm of the number of news documents shares on Twitter. 24

4.1 Overview of GisPy pipeline. $e_1, ..., e_n$ are contextual embeddings of tokens. . 29
4.2 Gist Inference Score formula 29
4.3 Individual indices of best Gist Inference Score (GIS) on *Reports (Low Gist) vs. Editorials (High Gist)*. All values are z-scores. 39
4.4 Individual indices of best GIS on *Methods (Low Gist) vs. Discussion (High Gist)*. All values are z-scores. 40
4.5 Individual indices of best GIS on *Disney Gist=No (Low Gist) vs. Gist=Yes (High Gist)*. All values are z-scores. 41

5.1 Example of input and output in causal relation extraction. 48
5.2 Architecture of TAKT 51
5.3 Reconstruction of the original figure of the semantic organization of the relations of efficient causality (extract) in Garcia's paper. 53
5.4 Overview of our proposed framework to continually pretrain PLMs with commonsense reasoning knowledge. 64
5.5 Distribution of samples in GeneraLized and COntextualized Story Explanations (GLUCOSE) and ATOMIC$^{20}_{20}$ based on the number of tokens (separated by space). 65
5.6 Distribution of relation types in ATOMIC$^{20}_{20}$. 66
5.7 Examples of converting two triples in ATOMIC$^{20}_{20}$ to natural language text using human-readable templates. 67
5.8 A sample that shows the annotated causal and temporal labels of an event pair. 69

5.9 Examples of Choice of Plausible Alternatives (COPA), Temporal and Causal Reasoning (TCR), and Penn Discourse Tree Bank 3.0 (PDTB3). 76

List of Tables

2.1 Results of classifying fake news vs. satire articles using Bidirectional Encoder Representations from Transformers (BERT) pretrained model. 12

2.2 Results of classifying fake news vs. satire articles using the Multinomial Naive Bayes baseline, Coh-Metrix features, and BERT. *Statistically significant difference with the baseline. 12

2.3 Significant components of our logistic regression model using the Coh-Metrix features. **Bold**: the remaining features following the step-wise backward elimination. 13

3.1 "RCs" are the Principal Component Analysis (PCA) components. "Description" contains the description of the Coh-Metrix index with the highest absolute loading value in the corresponding PCA component. 23

3.2 Results of running PCA. "RCs" are the PCA components. Significant components based on p-values in the linear regression model and factor loadings $\geq$ 0.40 are shown in the table. 25

3.3 Summary of the regression model of predicting truth labels. "RCs" are the PCA components. "Description" contains the description of the Coh-Metrix index with the highest absolute loading value in the corresponding PCA component. 26

4.1 Local and global index posfixes 30

4.2 Top 10 GIS scores computed for **Reports** (*Low Gist*) vs. **Editorials** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. 37

4.3 Top 10 GIS scores computed for **Methods** (*Low Gist*) vs. **Discussion** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. 37
4.4 Top 10 GIS scores computed by GisPy for **Gist=No** (*Low Gist*) vs. **Gist=Yes** (*High Gist*) articles in the Disney dataset. **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. 38
4.5 Comparison of GIS scores generated by GisPy vs. other baselines for all benchmarks. 38
4.6 GisPy GIS scores for train and test sets on all benchmarks. 40
4.7 Best combinations in robustness Test 2 on the train set for all experiments separated by benchmark. S/P Ratio: the number of sentences to the number of paragraphs. 42

5.1 Summary of some of the works on causal relation identification in natural language. 63
5.2 Example of an entry in GLUCOSE 68
5.3 List of features in CREST. All *value types* are Python values. 73
5.4 Results of adding COPA question types as Prompt vs. No Prompt on BERT-Large baseline. 75
5.5 TCR Accuracy results. 77
5.6 Penn Discourse Treebank Version 3.0 (PDTB3) Accuracy results. 77
5.7 COPA-test Accuracy results. 78
5.8 COPA-test Accuracy results on Easy and Hard question subsets. 79
5.9 BCOPA-CE Accuracy results. Base model in *b-l-** is BERT-Large. 80
5.10 Correctly classified samples in COPA paired with most semantically similar entries in GLUCOSE. 82

Abbreviations

BCOPA Balanced COPA 78

BERT Bidirectional Encoder Representations from Transformers xv, 10–13, 44, 49

CEA Cause-Effect Association 56, 58

CNN Convolutional Neural Network 62

COPA Choice of Plausible Alternatives xiv, 44, 73–76, 78, 81

CREST Causal RElation Schema for Text 6, 69, 71, 72

ESC Event StoryLine Corpus 60, 70, 71

FTT Fuzzy-trace theory 4, 5, 14, 16, 17, 28, 32

GCN Graph Convolutional Networks 62

GIS Gist Inference Score xi, xiii, xv, xvi, 28–30, 34–41, 43

GLUCOSE GeneraLized and COntextualized Story Explanations xiii, xvi, 44, 64, 65, 67, 68, 77, 80–82

HSI Hang Seng Index 53

ILP Integer Linear Programming 57, 76

KG Knowledge Graph 67, 78

KGs Knowledge Graphs 45, 64

LSA Latent Semantic Analysis 14, 18, 24, 32, 33

MLM Masked Language Modeling 44, 45, 64, 65, 68

NaCCS Narrative Coherence Coding System 20

NLG Natural Language Generation 58

NLP Natural Language Processing v, viii, 3, 5, 18, 20, 27–29, 49, 55, 61, 75

NLU Natural Language Understanding 3, 44, 49

PCA Principal Component Analysis xv, 10, 11, 22–26

PDTB Penn Discourse Treebank 56, 69, 70

PDTB3 Penn Discourse Tree Bank 3.0 xiv, 44, 73–76

PLM pretrained language model 44

PLMs pretrained language models 32, 33, 44, 45, 63, 77, 82

POS Part-of-speech 33, 34

RED Richer Event Description 70

SVM Support Vector Machine 12, 58

TCR Temporal and Causal Reasoning xiv, 44, 73–76

Chapter 1: Introduction

1.1 Motivation

The volume of digital data available globally has been growing at an unprecedented pace where nearly 80% of this data is estimated to be in the unstructured format (King, 2020). The human mind alone is not capable of efficiently mining and processing this unstructured data, which is often stored as text in natural language, at scale. Machines can process large amounts of data faster and more efficiently than humans can. However, by default, they lack an understanding of the text as we humans do and are not intelligent at the human level yet. This lack of intelligence has created the need to develop tools and algorithms for machines to understand the human language. These algorithms and methods are often studied and classified under the fields of Natural Language Processing (NLP) or Natural Language Understanding (NLU). One of the critical tasks in NLP is to identify the semantic relationships between various constituents of natural language that can facilitate many other tasks such as language modeling and recognizing textual entailment (Hendrickx, Kim, Kozareva, Nakov, Ó Séaghdha, et al., 2010). Among various semantic relationships, tools for the identification of causal relations are the ones that are not advanced while they can unlock massive value across various applications. Moreover, despite their importance in various applications, utilizing causal relations has been mainly limited to the core NLP tasks and did not find applications or implementations beyond these conventional tasks.

In this thesis, we initially demonstrate and investigate three impactful applications of causal relation identification algorithms: (1) identifying nuances between satire and fake news, (2) predicting the spread of information, including misinformation online, and (3) measuring gist inference score in text. These demonstrations additionally highlighted the need for developing more robust models to advance the identification of causal relations and demonstrated that efforts toward building such models are somewhat scattered, which

can hinder cross-collaborations and progress in this area. We address these limitations by building new language models augmented with commonsense knowledge to better detect causal relations. As part of this development, and motivated by the fact that causal relation schematics are somewhat scattered, we also introduce a new framework that can be used to unify a comprehensive set of data resources for causal relations in text to facilitate the development and evaluation of new models.

1.2 Thesis outline

Following this brief introduction and in Chapter 2, we demonstrate the utility of causal relations in developing methods of discerning between satire and false news. The blurry line between nefarious false news and protected-speech satire has been a significant struggle for social media platforms (Golbeck et al., 2018). To reduce and minimize users' exposure to misinformation on social media, purveyors of false news have begun masquerading as satire sites to avoid being systematically demoted. Contrary to false news, satire stories are usually humorous and may carry a political or social message and sometimes get mistakenly flagged as false content. To address this challenge, we investigate the linguistic and semantic cues between satire and false content to identify nuances. Furthermore, we show how causal relations and their proxies are among the significant predictors in distinguishing these groups.

Next, in Chapter 3, we introduce a framework for analyzing and predicting the spread of various types of information online. Fuzzy-trace theory (FTT) (Reyna, 2008), a leading account of decision making, suggests that humans encode stimuli, such as online content, at multiple levels of representation, namely gist, or bottom-line meaning, and verbatim, or surface-level details. Both of these levels of representation are expected to contribute independently to online information spread, with the effects of gist dominating. Essential aspects of gist in the context of online content include the presence of a clear *causal* structure and semantic coherence—both of which aid in meaning extraction. We test the hypothesis

that causal and semantic coherence are associated with the online sharing of misinformative social media content. Our results support FTT's predictions regarding the role of causally- and semantically-coherent content in promoting online sharing and motivate better measures of these critical constructs.

Next, in Chapter 4, and inspired by FTT, the utility of gist representations in human decision making and the lack of models for measuring gist in language, we develop computational tools for characterizing the capability of a given text to create a clear gist representation for readers. Most existing NLP tools and models focus on measuring coherence, cohesion, and readability in text. Even though coherence promotes gist extraction, these two concepts are not the same, and gist can be viewed as a mechanism that allows coherence apprehension (Glanemann et al., 2016). We build a new open-source tool in which causal relations are one of the major indices that can be used to automatically measure gist inferences for a collection of text documents. As part of this process, we also aggregate and standardize three benchmarks of low-gist and high-gist documents to facilitate the evaluation of computational tools for measuring gist in text.

Lastly, in Chapter 5, and following the demonstration of how and where causal relations play an important role in three impactful applications, we initially review the history of the automatic extraction of causal relations in text and highlight the need to increase the robustness of models for the accurate identification of these relations in natural language. Thereafter, and to address this need, we present a framework for building new models for the identification of causal relations in text. In this framework, we continually train pre-trained language models with commonsense reasoning knowledge hypothesizing that augmentation of such knowledge is helpful for the identification of cause-effect relations in text. During the evaluation of our framework, we also realized that efforts in creating causal relation benchmarks are somewhat scattered and do not often follow the same schematic or machine-readable format. Such an incompatibility among benchmarks has been slowing down the progress in the classification of semantic relations, including causal relations (Hendrickx,

Kim, Kozareva, Nakov, Séaghdha, et al., 2010). To address this problem, we also introduced the Causal RElation Schema for Text (CREST) to aggregate and unify an extensive collection of widely used resources of causal relation datasets and benchmarks to make them easy to access.

Throughout the thesis and in each chapter, we also comment on how our research can be used to guide the field through future developments of causal relation classification models and algorithms.

Chapter 2: Identifying Nuances Between Satire and Fake News

The blurry line between nefarious fake news and protected-speech satire has been a notorious struggle for social media platforms. Further to reducing exposure to misinformation on social media, purveyors of fake news have begun masquerading as satire sites to avoid being demoted. This chapter addresses the challenge of automatically classifying fake news versus satire. Previous work has studied whether fake news and satire can be distinguished based on language differences. Contrary to fake news, satire stories are usually humorous and carry political or social messages. We hypothesize that these nuances could be identified using semantic and linguistic cues that include proxies of *causal relations* in text, such as causal particles and connectives. Consequently, we train machine learning models using semantic representations computed with a state-of-the-art contextual language model and linguistic features based on textual coherence metrics. Empirical evaluation attests to the merits of our approach compared to the language-based baseline and sheds light on the nuances between fake news and satire.

2.1 Introduction

The efforts by social media platforms to reduce the exposure of users to misinformation have resulted, on several occasions, in flagging legitimate satire stories. To avoid penalizing publishers of satire, which is a protected form of speech, the platforms have begun to add more nuance to their flagging systems. Facebook, for instance, added an option to mark content items as "Satire" if "the content is posted by a page or domain that is a known satire publication, or a reasonable person would understand the content as irony or humor with a social message" (Facebook, 2019). This notion of humor and social message is also echoed in the definition of satire by the Oxford dictionary as "the use of humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, particularly in the context of

contemporary politics and other topical issues".

The distinction between fake news and satire carries implications with regard to the exposure of content on social media platforms. While fake news stories are algorithmically suppressed in the news feed, the satire label does not decrease the reach of such posts. This also affects the experience of users and publishers. For users, incorrectly classifying satire as fake news may deprive them of desirable entertainment content while identifying a fake news story as legitimate satire may expose them to misinformation. For publishers, the distribution of a story impacts their ability to monetize content. Moreover, in response to these efforts to demote misinformation, fake news purveyors have begun to masquerade as legitimate satire sites, for instance, carrying small badges at the footer of each page denoting the content as satire (Golbeck et al., 2018). The disclaimers are usually small such that the stories are still being spread as though they were real news (Funke, 2019). This gives rise to the challenge of classifying fake news versus satire based on the content of a story. While previous work has shown that satire and fake news can be distinguished with a word-based classification approach (Golbeck et al., 2018), our work is focused on the semantic and linguistic properties of the content. Inspired by the distinctive aspects of satire with regard to humor and social message, we hypothesize that using semantic and linguistic cues can help capture these nuances.

Our main research questions are; (1) are there semantic and linguistic differences between fake news and satire stories that can help to tell them apart?; (2) can these semantic and linguistic differences contribute to understanding nuances between fake news and satire beyond differences in the language being used? The rest of this chapter is organized as follows: in Section 2.2, we briefly review studies on fake news and satire articles that are the most relevant to our work. In Section 2.3, we present the methods we use to investigate semantic and linguistic differences between fake and satire articles. Next, we evaluate these methods and share insights on nuances between fake news and satire in Section 2.4. We

conclude the chapter in Section 2.5 and outline the next steps and future work.[1]

2.2 Related work

Previous work addressed the challenge of identifying fake news (Conroy et al., 2015; Shu, Sliva, et al., 2017) or identifying satire (Burfoot & Baldwin, 2009; Reganti et al., 2016; Rubin et al., 2016), in isolation, compared to real news stories.

The most relevant work to ours is that of Golbeck et al. (2018). They introduced a dataset of fake news and satirical articles, which we also employ in this work. The dataset includes the full text of 283 fake news stories and 203 satirical stories posted between January 2016 and October 2017, focusing mainly on American politics. These fake and satirical stories were verified manually such that each fake news article is paired with a rebutting article from a reliable source. This data carries two desirable properties. First, the labeling is based on the content, not the source, and stories are spread across a diverse set of sources. Second, as also mentioned by Golbeck et al. (2018), the fact that fake news and satire articles both focus on American politics minimizes the possibility that the topic of the articles will influence the classification.

In their work, Golbeck et al. (2018) studied whether there are differences in the language of fake news and satirical articles on the same topic that could be utilized with a word-based classification approach. Their paper proposes a model using the Multinomial Naive Bayes algorithm, which serves as the baseline in our experiments.

2.3 Method

In the following subsections, we investigate the semantic and linguistic differences of satire and fake news articles.[2]

[1] This chapter contains a discussion of the work published in (Levi, Hosseini, Diab, & Broniatowski, 2019)
[2] Codes and reproducibility reports available at: https://github.com/adverifai/Satire_vs_Fake

2.3.1 Semantic representation with BERT

To study the semantic nuances between fake news and satire and compute the vector representation of words in context, we leverage BERT (Devlin et al., 2018), a contextual language model. Even though there are other tools such as Word2Vec (Mikolov et al., 2013) to produce word vectors that capture words' semantic or meaning-related relationships, such tools are mainly context-free models that generate a single word-embedding for each word in the vocabulary. BERT, on the other hand, generates a representation of each word based on the other words in context. BERT was built upon recent work in pretraining contextual representations, such as ELMo (Peters et al., 2018) and ULMFiT (Howard & Ruder, 2018), and is deeply bidirectional, representing each word using both its left and right context. We use the pretrained models of BERT and fine-tune them on the dataset of fake news and satire articles. Our BERT-based binary classifier is created by adding a single new layer to BERT's neural network architecture that will be trained to fine-tune BERT to our task of classifying fake news and satire articles.

2.3.2 Linguistic analysis with Coh-Metrix

Inspired by previous work on satire detection, specifically, Rubin et al. (2016) who studied the humor and absurdity aspects of satire by comparing the final sentence of a story to the first one, and to the rest of the story, we hypothesize that metrics of text coherence and cohesion will be useful to capture similar aspects of semantic relatedness between different sentences of a story. To compute these text coherence metrics, we leverage Coh-Metrix (McNamara et al., 2010). Coh-Metrix is a tool for producing linguistic and discourse representations of a text. As a result of applying the Coh-Metrix to the input documents, we have 108 indices related to text statistics, such as the number of words and sentences; referential cohesion, which refers to overlap in content words between sentences; various text readability formulas; different types of connective words and more. To account for multicollinearity among the different features, we first run a PCA on the set of Coh-Metrix

indices. It is worth noting that we do not apply dimensionality reduction, and features still correspond to the Coh-Metrix indices and are thus explainable. After running the PCA, we use the PCA scores as independent variables in a logistic regression model with the fake and satire labels as our dependent variable. Afterwards, we also run a step-wise backward elimination regression to see what components are also significant in the step-wise model.

2.4 Results and discussion

In the following subsections, we evaluate our classification model and share insights on the nuances between fake news and satire, while addressing our two research questions.

2.4.1 Classification between fake news and satire

We evaluated the performance of our method using a dataset of fake news and satire articles and the F1 score with 10-fold cross-validation, as was done in the previous work (Golbeck et al., 2018).

First, we considered the semantic representation with BERT. Our experiments included multiple pretrained models of BERT with different sizes and case sensitivity, among which the large uncased model, `bert_uncased_L-24_H-1024_A-16`, achieved the best results. We use the recommended settings of hyper-parameters in BERT's GitHub [3] repository. Furthermore, we tested separate models based on the headline and body text of a story, and in combination. Results are shown in Table 2.1. The models based on the headline and text body give a similar F1 score. However, while the headline model performs poorly on precision, perhaps due to the short text, the model based on the text body performs poorly on recall. The model based on the full text of headline and body gives the best performance.

To investigate the predictive power of the linguistic cues, we use the Coh-Metrix indices that were significant in both the logistic and step-wise backward elimination regression models, and train a classifier on fake news and satire articles. We tested a few classification

[3] https://github.com/google-research/bert

Model	Precision	Recall	F1
Headline only	0.46	**0.89**	0.61
Text body only	0.78	0.52	0.62
Headline + text body	**0.81**	0.75	**0.78**

Table 2.1: Results of classifying fake news vs. satire articles using BERT pretrained model.

Method	Precision	Recall	F1
Baseline	0.70	0.64	0.67
Coh-Metrix	0.73	**0.76**	0.77*
Fine-tuned BERT	**0.81**	0.75	**0.78***

Table 2.2: Results of classifying fake news vs. satire articles using the Multinomial Naive Bayes baseline, Coh-Metrix features, and BERT. *Statistically significant difference with the baseline.

models, including Naive Bayes, Support Vector Machine (SVM), logistic regression, and gradient boosting—among which the SVM classifier achieved the best results.

Table 2.2 provides a summary of the results. We compare the results of our methods of the pretrained BERT, using both the headline and text body, and the Coh-Metrix approach, to the language-based baseline with Multinomial Naive Bayes from (Golbeck et al., 2018).[4] For Coh-Metrix, we report the *mean* Precision, Recall, and F1 on the test set. Both the semantic cues with BERT and the linguistic cues with Coh-Metrix significantly outperform the baseline on the F1 score. The two-tailed paired t-test with a 0.05 significance level was used for testing the statistical significance of performance differences. The best result is given by the BERT model. Overall, these results answer research question 1 regarding the existence of a semantic and linguistic difference between fake news and satire.

[4]We were not able to reproduce the same result as in the original paper, possibly due to the difference between scikit-learn and Weka, tools that were used to train classifiers in our work and the original paper, respectively. To make our results comparable, we replicated the experiments of the original paper using scikit-learn.

2.4.2 Insights on linguistic nuances

Significant features of the logistic regression model we built using Coh-Metrix indices are shown in Table 2.3 with the respective significance levels. Those components that are also significant in the step-wise backward elimination regression appear in bold.

	PCA Component	Description	Estimate	std.error	t-statistic	
Satire associated	**RC19**	First person singular pronoun incidence	1.80	0.41	4.38	***
	RC5	Sentence length, number of words	0.66	0.18	3.68	***
	RC15	Estimates of hypernymy for nouns	0.61	0.19	3.18	**
	RC49	Word Concreteness	0.54	0.17	3.18	**
	RC35	Ratio of causal particles to causal verbs	0.56	0.18	3.10	**
	RC91	Text Easability PC Referential cohesion	0.45	0.16	2.89	**
	RC20	Incidence score of gerunds	0.43	0.16	2.77	**
	RC32	Expanded temporal connectives incidence	0.44	0.16	2.75	**
	RC9	Third person singular pronoun incidence	0.44	0.16	2.67	**
	RC43	Word length, number of letters	0.45	0.20	2.27	*
	RC46	Verb phrase density	0.37	0.16	2.25	*
	RC97	Coh-Metrix L2 Readability	0.34	0.16	2.16	*
	RC61	Average word frequency for all words	0.50	0.24	2.13	*
	RC84	The average givenness of each sentence	0.37	0.18	2.11	*
	RC65	Text Easability PC Syntactic simplicity	0.38	0.18	2.08	*
	RC50	Lexical diversity	0.37	0.18	2.05	*
Fake news associated	**RC30**	Agentless passive voice density	-1.05	0.21	-4.96	***
	RC73	Average word frequency for content words	-0.72	0.20	-3.68	***
	RC59	Adverb incidence	-0.62	0.18	-3.43	***
	RC55	Number of sentences	-0.79	0.26	-3.09	**
	RC62	Causal and intentional connectives	-0.42	0.15	-2.72	**
	RC34	LSA overlap between verbs	-0.35	0.16	-2.22	*
	RC44	LSA overlap, adjacent sentences	-0.36	0.16	-2.16	*
	RC47	Sentence length, number of words	-0.36	0.18	-2.03	*
	RC89	LSA overlap, all sentences in paragraph	-0.34	0.17	-1.97	*
	(Intercept)		-0.54	0.19	-2.91	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.3: Significant components of our logistic regression model using the Coh-Metrix features. **Bold**: the remaining features following the step-wise backward elimination.

With regard to research question 2 on the understanding of semantic and linguistic nuances between fake news and satire, a key advantage of studying the coherence metrics is explainability. While BERT gives us the best result, it is not easily interpretable. The coherence metrics allow us to study the differences between fake news and satire in a straightforward manner.

By observing the significant features, in bold in Table 2.3, we see a combination of surface level related features, such as sentence length and average word frequency, as well as semantic features including Latent Semantic Analysis (LSA) overlaps between verbs and between adjacent sentences. Semantic features which are associated with the gist representation of content are particularly interesting to see among the predictors because based on Fuzzy-trace theory (FTT) (Reyna, 2012a), a well-known theory of decision making under risk, gist representation of content drives individual's decision to spread misinformation online. Also among the significant features, we observe the *causal connectives*, that are proven to be important in text comprehension, and two indices related to the text easability and readability, both suggesting that satire articles are more sophisticated, or less easy to read, than fake news articles.

2.5 Conclusion and future work

We addressed the challenge of identifying nuances between fake news and satire. Inspired by the humor and social message aspects of satire articles, we tested two classification approaches based on a state-of-the-art contextual language model and linguistic features of textual coherence. Evaluation of our methods pointed to the existence of semantic and linguistic differences between fake news and satire. In particular, both methods achieved a significantly better performance than the baseline. Lastly, we studied the feature importance of our linguistic-based method to help shed light on the nuances between fake news and satire. We observed that satire articles are more sophisticated, or less easy to read, than fake news articles.

Overall, our contributions, with the improved classification accuracy and the understanding of nuances between fake news and satire, carry significant implications with regard to the delicate balance of fighting misinformation while protecting free speech. For future work, we plan to study additional linguistic cues, specifically humor-related features, such as absurdity and incongruity, which were shown to be good indicators of satire in previous

work. Another interesting line of research would be to investigate techniques of identifying whether a story carries a political or social message, for example, by comparing it with timely news information.

Chapter 3: Does Causal Coherence Predict Online Spread of Social Media?

Most people who propagate false information online do so on purpose. We therefore seek to understand the factors making this content compelling and, ultimately, driving online sharing. Fuzzy-trace theory (FTT) (Reyna, 2012a), a leading account of decision making, posits that humans encode stimuli, such as online content, at multiple levels of representation; namely, gist, or bottom-line meaning, and verbatim, or surface-level details. Both of these levels of representation are expected to contribute independently to online information spread, with the effects of gist dominating. Important aspects of gist in the context of online content include the presence of a clear causal structure and semantic coherence—both of which aid in meaning extraction. In this chapter, we test the hypothesis that *causal* and semantic coherence are associated with online sharing of misinformative social media content using Coh-Metrix—a widely-used set of psycholinguistic measures. Results support FTT's predictions regarding the role of causally- and semantically-coherent content in promoting online sharing and motivating better measures of these key constructs.

3.1 Introduction

The speed and scale with which false news spreads through online platforms have posed new challenges to society, with implications for politics, the economy, and mental health. Although automated accounts ("bots") have been implicated in dissemination of such misinformation—especially at early stages on social media networks (Shao et al., 2018)—recent studies demonstrate that human users are the primary vector for the spread of false content (Vosoughi et al., 2018). We therefore seek to understand what content features lead humans to spread false information. Despite the numerous studies focusing on meta-information, such as social network structure (Shu et al., 2019), individual personality differences (Pennycook & Rand, 2018), source credibility (Grinberg et al., 2019), and prior

exposure (Pennycook et al., 2018), there exists relatively little focus on online *content*.

We seek to determine what characteristics of online content make this content compelling, thereby encouraging people to share it. Our analysis is based on Fuzzy-trace theory (FTT) (Reyna, 2012a), a leading theory of decision-making under risk. According to FTT, there are multiple representations of online content encoded in parallel: (1) the gist representation is a developmentally-advanced, subjective, simplified, memorable, and contextualized representation of the bottom-line meaning of a stimulus, and (2) the verbatim representation is detailed, factual, brittle in memory, and decontextualized. In the context of online misinformation, gist tends to be a semantically- and causally-coherent representation of the situation described in the online content, typically linking some events, actors, and outcomes —regardless of its truth value (Broniatowski & Reyna, 2017; Reyna, 2012b).

In practice, humans tend to rely on gist representations when making decisions, such as whether or not to share online content. However, prior work examining the role of content in online information spread has primarily focused on verbatim content features, e.g., (Tsur & Rappoport, 2012), although see (Dredze et al., 2016).

In this chapter, we will test the hypothesis that more causally- and semantically- coherent text documents —both necessary (but not sufficient) conditions for meaningful gist— are more likely to be shared by individuals on social media. We will briefly review the literature on semantic and causal coherence measures and mental representation of text content. Based on this review, we will use indices derived from Coh-Metrix (McNamara et al., 2010) to compute various indices, including causality indices in text, and run a linear regression analysis using these indices. In particular, the chapter aims to investigate the associations between Coh-Metrix indices related to causal and semantic coherence and the number of shares of a collection of news documents on Twitter.

The rest of this chapter is organized as follows: Section 3.2 will review the existing metrics of causal and semantic coherence, mainly from the psychology literature. Section 3.3 describes the analysis of a dataset of fake and real news articles to test the hypothesis. The

next Section 3.4 will discuss the results of linear regression analysis. Finally, we will conclude the chapter in Section 3.5 and outline the next steps and future work.[1]

3.2 Related work

This section reviews the methods of measuring causal and semantic coherence in text documents. It is important to emphasize that semantic coherence in text documents has been measured using different techniques, including topic models (Mimno et al., 2011) and Latent Semantic Analysis (LSA) (Foltz et al., 1998). However, coherence has different types, and our intention here is to measure ***causal*** *coherence*. Since there exists a literature gap in studies measuring causal coherence using Natural Language Processing (NLP) methods, this study will mainly review the psychology and psycholinguistics literature to find a metric for causal coherence that can be implemented using text analysis techniques.

At first glance, most of the methods of rating causal coherence are based on discovering *causal links* among *units* of information in content. As a result, there are two key questions to address when finding a metric for causal coherence. First, what is a "unit" of information in a document? A unit of information could be in the form of a sentence, clause, event, action, state, or any combination of them. Second, how do we define and extract causal links between these information units? In the following, we address these two questions.

In most studies, the definition of causal coherence is based on causal links in documents. In (Black & Bern, 1981), causal coherence is measured by checking if there is a causal link between a pair of sentences. Here, causal relations were manually extracted. The definition of causal coherence in (Kuperberg et al., 2011) is also based on the existence of causal connections across sentences, which participants manually check by assigning a score between 1 to 7 to triplets of sentences. In (Trabasso et al., 1982) there is no definition for causal coherence; however, causal cohesion for a story is defined in terms of the percentages of events in the story that are contained in the causal chain and as the percentage of events

[1]This chapter contains a discussion of the work published in (Hosseini, Diab, & Broniatowski, 2019)

with causal chain connections. In (Mulder & Sanders, 2012), the question regarding the unit of information in documents is addressed. A causal coherence relation, explicit or implicit, is not a relation between propositions. Rather, coherence between two sentences describing cause and effect, respectively, is the result of the fact that *events* described in these sentences are causally related in the world described by the text —not between the *sentences*. Thus, inferring causal coherence requires background knowledge.

Grysman and Hudson (2010) introduced three measures of causal coherence. These measures include (1) meaning-making, (2) narrative complexity, and (3) the use of causal terms. Meaning-making refers to the tendency of subjects to describe lessons or insights that they gained connected to the story they told. Narrative complexity analyzes the recognition of multiple dimensions, perspectives, or emotions in a narrative, and the coding scheme measures narrative complexity. For measuring causal terms, fifteen causal terms were coded. The frequency of word occurrence was obtained using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) software, which reports the frequency of occurrence of categories of words in narrative units.

A summary of what we need to know about causal coherence and how it aids readers' comprehension of narratives is provided by Arfé and Boscolo (2006). Authors refer to Trabasso et al. (1982)'s definition of causal coherence in their work: "causal coherence is a property of stories that readers mentally represent as a causal network, where events are interconnected through causal links." One interesting point mentioned here is that causal coherence is not strictly dependent on elaborate syntactic structures or cohesion devices (such as causal conjunctions) but depends mostly on conceptual links of causality. Their unit of analysis in the text is the "T-unit," or terminable unit, which refers to the main clause and any subordinate clauses attached to it. Finally, they distinguish between different causal relation types (enabling, physical, psychological, and motivational) —with a focus on the relations particularly relevant to the narration of human-like events: psychological and

motivational— and between local vs. global coherence. [2]

Diehl et al. (2006) use Trabasso and Sperry (1985)'s paradigm to identify causal connections and a story's causal chain. Here, a causal connection refers to the direct causal relationship between two story events, which can be evaluated with a necessity test (i.e., event X is caused by event Y if event X would NOT have occurred if event Y had not occurred).

In (Reese et al., 2011) a scale for coding coherence called the Narrative Coherence Coding System (NaCCS) is used. NaCCS has three dimensions including Context, Chronology, and Theme. In the Theme dimension (on the scale of 0-3) which is for assessing the meaning-making aspect of a narrative, the existence of causal linkages is checked. Specifically, they mention in their criteria that a narrative that has an identifiable topic but in which the topic is not developed via elaborations, evaluations, and causal linkages receives a 1. And to receive a score of 2 on this dimension, the narrative must substantially develop the topic via causal linkages, elaborations, evaluations, and interpretations.

In summary, measuring causal coherence is highly dependent on the **causal connections** between **events** in a text document. When we extract the causal relations between every pair of events in a document, we will be able to measure the causal coherence of the document in a variety of ways. For instance, we can build a causal chain and assess the level of causal coherence in a document based on the length of the causal chain. Or, we can build a causal network based on Trabasso's model, where nodes are the events in the document and edges are the causal links between events and define a causal coherence score using such a network.

Extracting implicit causal relations is a challenging open task in NLP. As mentioned in (Arfé & Boscolo, 2006), causal coherence is not strictly dependent on the use of elaborate syntactic structures or cohesion devices (such as causal conjunctions) but depends mostly on conceptual links of causality, which are mostly implicit in content. Even though there is

[2]Finding local and global coherence is based on the work by Trabasso and Sperry (1985).

still no publicly available tool for tagging both implicit and explicit causal relations in text with high and reliable accuracy, we can utilize tools that at least provide us with an estimate of the degree to which a document contains causal relations.

3.3 Method and analysis

In this section, we test [3] our hypothesis that more causally coherent documents are more likely to be shared on social media.

3.3.1 Dataset and preprocessing

We use FakeNewsNet (Shu, Wang, & Liu, 2017), a dataset of fake and real news articles and their social context information. We chose this dataset because it contains an article's content (body text) and measures of social engagement and the number of shares on Twitter. Furthermore, these articles have been fact-checked, enabling us to know their true value.

We filtered the dataset and removed news documents that did not have a proper value in their *Body* feature. In particular, we removed those records that either had no text or included only one sentence, potentially due to unsuccessful parsing of the corresponding HTML files. After the filtering process, 414 news articles remained in the dataset.[4] We did *not* remove punctuation marks and stop words. In addition, we did not do any stemming, lemmatizing, or lower-casing to avoid any possible impact on Coh-Metrix's output. All the paragraphs in input text documents are separated by new line characters.

3.3.2 Measuring causal coherence

We use Coh-Metrix (McNamara et al., 2010) to compute a suite of features for coherence and cohesion of text documents. Cohesion is defined as characteristics of the explicit text that play a role in assisting the reader in mentally connecting ideas in the text (Graesser

[3] Codes and reproducibility reports available at: https://github.com/phosseini/SBP-BRiMS2019
[4] Our inputs to the Coh-Metrix are all plain text stored in .txt file format.

et al., 2003). Coh-Metrix is a tool for producing the indices of linguistic and discourse representation of a text. Coh-Metrix indices can be used to investigate the cohesion of the text and the ease with which a coherent underlying mental representation may be generated. There are 108 different indices produced by Coh-Metrix which lie in several categories pertinent to text comprehension. Among these indices, we are looking for those which may have some association with causal and semantic coherence constructs.

3.3.3 Analysis

Coh-Metrix yielded 108 indices for each document; however, several of these features were correlated. To account for multicollinearity, we applied PCA on this set of features before doing our regression analysis. PCA was conducted with *principal* package in R.

After applying a logarithm transform to the dependent variable (distinct number of shares) to control for skew,[5] and using the PCA factors, truth (fake/real) labels, and source of fact-checking as inputs, we conducted a multiple linear regression analysis to determine which factors predicted online sharing.

3.4 Results and discussion

Results of the regression analysis, shown in Table 3.1, indicate that there are 16 significant predictors. Of these, one is a measure of referential cohesion. Referential cohesion aids readers in making connections between propositions, clauses, and sentences, enabling semantic coherence and helping readers to create a coherent mental representation of text. A greater cohesion indicates greater coherence in mental representation and gist comprehension (Wolfe et al., 2018). As a result, presence of referential cohesion as one of the significant variables in our linear regression model demonstrates the importance of gist in predicting online sharing.

Also, as shown in Table 3.2, the *RC2* component is highly associated with the semantic

[5]Normal probability plots show that residuals follow a normal distribution, see Figure 3.1.

Components	Description	Estimate	std.error	t-statistic	
RC58	Number of syllables	0.11	0.03	3.56	***
Truth labels	Fake/Real labels	0.32	0.09	3.54	***
RC14	Preposition phrase density	-0.10	0.03	-3.27	**
RC16	Adverbial phrase density	-0.09	0.03	-3.02	**
RC2	Argument overlap, all sentences	0.09	0.03	2.86	**
RC24	First person plural pronoun incidence	0.09	0.03	2.80	**
RC72	Referential cohesion	-0.07	0.03	-2.42	*
RC7	Number of words	0.08	0.03	2.41	*
RC11	Number of letters	-0.07	0.03	-2.33	*
RC47	LSA overlap, adjacent paragraphs	-0.07	0.03	-2.22	*
RC31	Third person plural pronoun	-0.07	0.03	-2.18	*
RC66	Stem overlap, adjacent sentences	0.07	0.03	2.14	*
RC94	Hypernymy for nouns and verbs	-0.06	0.03	-2.05	*
RC30	Temporal connectives incidence	0.06	0.03	2.02	*
RC39	Word frequency for content words	-0.06	0.03	-2.00	*
RC37	Noun phrase density	-0.06	0.03	-1.98	*
(Intercept)	-	1.52	0.13	11.95	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.1: "RCs" are the PCA components. "Description" contains the description of the Coh-Metrix index with the highest absolute loading value in the corresponding PCA component.

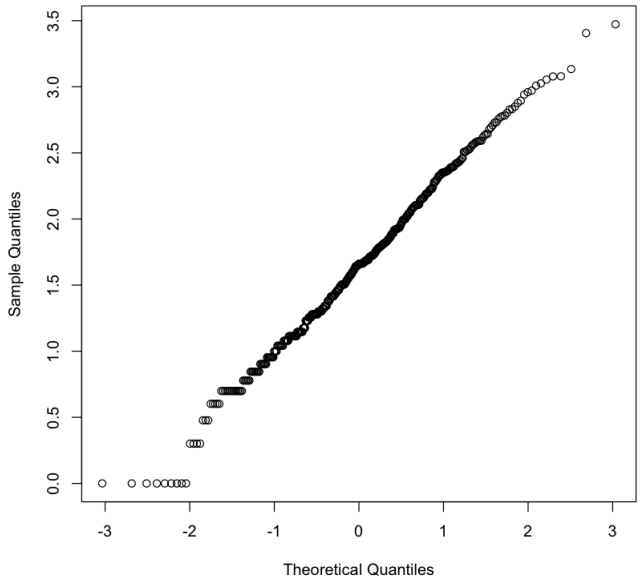


Figure 3.1: Normal probability of logarithm of the number of news documents shares on Twitter.

features of text. The LSA related features measure the semantic overlap between sentences or paragraphs.

Another significant predictor in our linear regression model is news documents' truth labels (fake/real). We conducted another experiment using the PCA scores in a stepwise backward elimination regression with the truth label as the dependent variable to understand better what features predict and explain these truth labels. A summary of the regression model is shown in Table 3.3. There are 32 significant components in the model. Based on the loading values of the PCA, there are three Coh-Metrix indexes, PCDCz, PCDCp, and CNCCaus, highly associated with one of the significant predictors, *RC5*. PCDCz and PCDCp which are indexes of deep cohesion demonstrate the degree to which the text contains causal and intentional connectives. According to the documentation of Coh-Metrix, these connectives help the reader form a more coherent and deeper understating of causal events in the text. Also, CNCCaus indicates the incidence score of causal connectives in documents. As a result, causal and deep cohesion connectives predict truth assessment (label), which is a significant predictor of online sharing.

Coh-Metrix Index	PCA Component										
	RC2	**RC7**	**RC16**	**RC47**	**RC37**	**RC30**	**RC24**	**RC11**	**RC31**	**RC14**	**RC58**
PCREFz	0.76										
PCREFp	0.74										
CRFNO1	0.89										
CRFAO1	0.87										
CRFSO1	0.85										
CRFNOa	0.88										
CRFAOa	0.91										
CRFSOa	0.85										
CRFCWO1	0.64										
CRFCWOa	0.69										
LSASSp	0.53										
LSAPP1	0.44										
LSASS1	0.66										
LSAGN	0.47										
DESPC		0.91									
DESSC		0.95									
DESWC		0.96									
LDTTRc		-0.61									
LDTTRa		-0.66									
DRAP			0.84								
WRDADV			0.66								
LSAPP1d				0.91							
DRNP					-0.91						
CNCTemp						0.94					
WRDPRP1p							0.96				
DESWLsyd								0.49			0.40
DESWLltd								0.80			
WRDPRP3p									0.91		
DRPP										-0.81	

Table 3.2: Results of running PCA. "RCs" are the PCA components. Significant components based on p-values in the linear regression model and factor loadings ≥ 0.40 are shown in the table. A detailed explanation of Coh-Metrix features is available on Coh-Metrix's website at: http://www.cohmetrix.com/.

3.5 Conclusion and future work

In the first part of this chapter, using Coh-Metrix, we tested our hypothesis that documents with more semantic and causal coherence are more likely to be shared on social media by individuals. Our results show statistically significant associations between proxies of these constructs and the number of shares of news documents.

The deep cohesion score in Coh-Metrix is mainly based on the incidence of explicit causal connectives. Thus, there is reason to believe that this metric could be improved. For future work, it would be interesting to extract the *implicit* causal relations from text

PCA Component	Description	Sum sq	Mean sq	t-statistic	
RC1	Number of letters	1.09	1.09	8.05	**
RC2	Argument overlap, all sentences	1.46	1.46	10.73	**
RC5	Text Easability PC Deep cohesion	1.05	1.05	7.77	**
RC7	Number of words	3.61	3.61	26.61	***
RC9	Additive connectives incidence	0.55	0.55	4.08	*
RC3	Sentence syntax similarity	0.71	0.71	5.22	*
RC15	Negative connectives incidence	0.89	0.89	6.55	*
RC17	Number of sentences	0.87	0.87	6.41	*
RC16	Adverbial phrase density	0.85	0.85	6.25	*
RC8	Number of words	1.75	1.75	12.88	***
RC27	Infinitive density	2.21	2.21	16.26	***
RC12	Word frequency in sentences	5.71	5.71	42.07	***
RC37	Noun phrase density	1.65	1.65	12.18	***
RC30	Temporal connectives incidence	0.60	0.60	4.44	*
RC31	Third person plural pronoun	2.07	2.07	15.28	***
RC19	Agentless passive voice density	0.62	0.62	4.56	*
RC89	Lexical diversity, type-token ratio	2.03	2.03	14.98	***
RC18	Text Easability PC Temporality	0.65	0.65	4.80	*
RC46	Age of acquisition for content words	1.77	1.77	13.06	***
RC60	Anaphor overlap, all sentences	2.22	2.22	16.37	***
RC39	Word frequency for content words	0.86	0.86	6.34	*
RC58	Number of syllables	1.36	1.36	10.02	**
RC50	Sentence syntax similarity	2.08	2.08	15.31	***
RC75	Number of paragraphs	0.70	0.70	5.16	*
RC79	Content word overlap, adjacent sentences	1.99	1.99	14.69	***
RC67	Number of letters	0.61	0.61	4.48	*
RC87	Text Easability PC Word concreteness	1.06	1.06	7.83	**
RC84	Text Easability PC Deep cohesion	0.65	0.65	4.80	*
RC86	Concreteness for content words	1.34	1.34	9.85	**
RC91	Stem overlap, all sentences	1.50	1.50	11.03	***
RC92	Positive connectives incidence	0.67	0.67	4.96	*
RC62	Noun overlap, all sentences	0.56	0.56	4.16	*

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.3: Summary of the regression model of predicting truth labels. "RCs" are the PCA components. "Description" contains the description of the Coh-Metrix index with the highest absolute loading value in the corresponding PCA component.

and build a causal network to better model readers' mental representation. Tagging causal relations in documents, specifically those of implicit type, is not an easy task in NLP. We believe that the focus in automatically modeling causal gist and mental representation of readers should be on extracting causal relations. Because there is a fairly clear idea of how causal coherence can be modeled from a psychological perspective, having a reliable tool for extracting events and causal relations between them in the content will enable us to better model causal coherence and, consequently, gist. This will enable us to test any hypothesis on the relation between this model and patterns of information sharing online.

Chapter 4: Measuring Gist Inference Score in Text

According to Fuzzy-trace theory (FTT) (Reyna, 2008, 2012a), when individuals read text, they encode multiple mental representations of the text parallel to their minds. These mental representations vary along a continuum ranging from *gist* to *verbatim*. While verbatim representations are related to surface-level information, gist represents the bottom-line meaning of the text, given its context. FTT interprets the word gist in roughly the same manner as common usage, namely, as the essence or central portion, the substance or pith of a matter. It is crucial to measure gist representations because they impact judgments and decisions more than verbatim representations (Reyna, 2021). The gist knowledge helps us measure the capability of a document (e.g., news article, social media post, etc.) in creating a clear and actionable mental representation in readers' minds and the degree to which a document can communicate its message.

The majority of existing NLP tools and models focus on measuring coherence, cohesion, and readability in text (Graesser et al., 2004; Lapata et al., 2005; Lin et al., 2011; Crossley et al., 2016; S. Liu et al., 2020; Laban et al., 2021; Duari & Bhatnagar, 2021). Even though coherence promotes gist extraction, these two are not the same. And gist can be viewed as a mechanism that allows coherence apprehension (Glanemann et al., 2016). To the best of our knowledge, there is no publicly available tool for directly measuring gist in text. Wolfe, Dandignac, and Reyna (2019), Dandignac and Wolfe (2020), and Wolfe, Dandignac, Wang, and Lowe (2021) are the only studies that introduced a theoretically motivated method to measure Gist Inference Score (GIS) using Coh-Metrix (Graesser et al., 2004), a tool for producing linguistic and discourse representations of a text including measures of cohesion and readability. Coh-Metrix, even though useful and inspiring, has several limitations. For instance, its public version does not provide batch processing of documents, is only accessible through a web interface, and its cohesion indices are centered

on local and global cohesiveness (Crossley et al., 2016). In this work,[1] inspired by (Wolfe et al., 2019) and definition of a subset of indices in Coh-Metrix, we build a new open-source tool, GisPy,[2] to automatically compute GIS for a collection of text documents. We leverage the state-of-the-art NLP tools and models such as contextual language model embeddings to further improve the quality of indices in our tool.

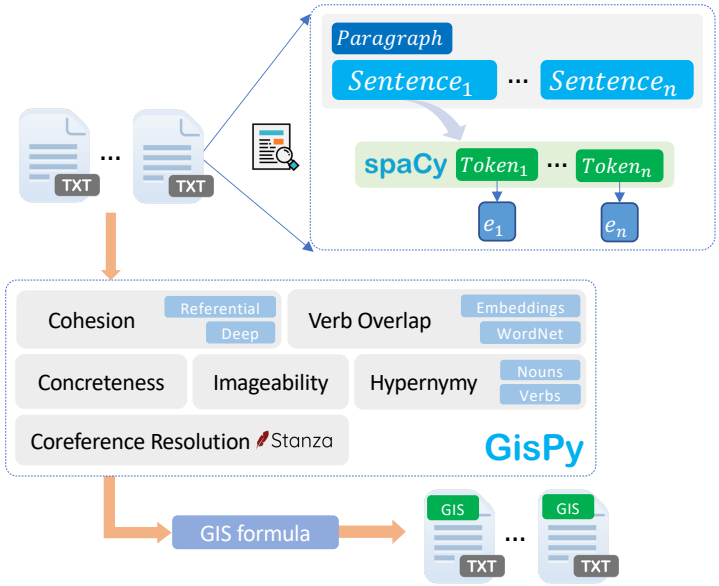


Figure 4.1: Overview of GisPy pipeline. $e_1, ..., e_n$ are contextual embeddings of tokens.

Figure 4.2: Gist Inference Score formula by Wolfe et al. (2019)

Our contributions can be summarized as follows:

- We introduce the first open-source and publicly available tool to measure GIS in text.
- We unify and standardize three benchmarks for measuring gist in text and report improved baselines on these benchmarks.

[1]This chapter contains a discussion of the work published in (Hosseini, Wolfe, et al., 2022)
[2]Our tool is publicly available to use at: https://github.com/phosseini/GisPy.

- By leveraging the explainability of indices in our tool, we investigate the role of individual indices in producing GIS for low vs. high gist documents across benchmarks.

4.1 Methods

This section explains how we implement each of the indices in GisPy and compute GIS. We start by explaining common implementation features among indices, followed by their specific details.

4.1.1 Local vs. global indices

We have taken different approaches in implementing indices for which we need to compute the overlap between words or sentences (e.g., semantic similarity). In particular, these indices are computed in two settings: (1) *local* and (2) *global*. In the local setting, we only take into account *consecutive/adjacent* words/sentences whereas in the global setting, we consider *all* pairs not just consecutive ones. Moreover, we compute indices one time by separating the paragraphs in text and another time by disregarding the paragraph boundaries. For clarity, we use postfixes listed in Table 4.1 for these variations.

Postfix	Explanation
$*_1$	Local index ignoring paragraph boundary
$*_a$	Global index ignoring paragraph boundary
$*_1p$	Local index at paragraph-level
$*_ap$	Global index at paragraph-level

Table 4.1: Local and global index posfixes

We assume every document is broken into paragraphs $\{P_0, P_1, ..., P_n\}$, separated by at least one newline character, each with one or more sentences $\{S_{0,0}, S_{0,1}, ..., S_{i,j}\}$ where each sentence has one or more tokens $\{t_{0,0,0}, t_{0,0,1}, ..., t_{i,j,k}\}$. As an illustration, consider the following for a document with two paragraphs each with two and three sentences,

respectively:

$$P_0 \rightarrow \{S_{0,0}, S_{0,1}\}$$

$$P_1 \rightarrow \{S_{1,0}, S_{1,1}, S_{1,2}\}$$

Where $S_{i,j}$ is the j_{th} sentence of paragraph i, this is how we compute local and global versions of index X —assuming X measures the similarity among sentences and similarity is computed by $\oplus$:

$$\begin{aligned}
X_1 = mean(&S_{0,0} \oplus S_{0,1}, S_{0,1} \oplus S_{1,0},\\
&S_{1,0} \oplus S_{1,1}, S_{1,1} \oplus S_{1,2})\\
X_a = mean(&S_{0,0} \oplus S_{0,1}, S_{0,0} \oplus S_{1,0},\\
&S_{0,0} \oplus S_{1,1}, S_{0,0} \oplus S_{1,2},\\
&S_{0,1} \oplus S_{1,0}, S_{0,1} \oplus S_{1,1},\\
&S_{0,1} \oplus S_{1,2}, S_{1,0} \oplus S_{1,1},\\
&S_{1,0} \oplus S_{1,2}, S_{1,1} \oplus S_{1,2})\\
X_1p = mean(&S_{0,0} \oplus S_{0,1}, S_{1,0} \oplus S_{1,1},\\
&S_{1,1} \oplus S_{1,2})\\
X_ap = mean(&S_{0,0} \oplus S_{0,1}, S_{1,0} \oplus S_{1,1},\\
&S_{1,0} \oplus S_{1,2}, S_{1,1} \oplus S_{1,2})
\end{aligned}$$

4.1.2 GisPy indices implementation

Referential Cohesion: This index (`PCREFz` in Coh-Metrix) reflects the overlap of words and ideas across sentences and the entire text. To measure this overlap, we leverage the

`Sentence Transformers` (Reimers & Gurevych, 2019) [3] to compute the embeddings of all sentences in a document using the `all-mpnet-base-v2` model.[4] We chose this model since it provides the best quality and has the highest average performance among all the other models introduced by Reimers and Gurevych (2019). Once we computed the embeddings, to measure the overlap across all sentences, we found the cosine similarity between embeddings of every pair of sentences, one time at paragraph level and another time ignoring the paragraph boundaries. This process results in four indices of referential cohesion including: `PCREF_1, PCREF_a, PCREF_1p, PCREF_ap`.

We additionally implement a new index based on coreference resolution in paragraphs in a document. In particular, using Stanford CoreNLP's coreference tagger (Manning et al., 2014) through Stanza's wrapper (Qi et al., 2020), we first find the number of coreference chains (`corefChain`) to the number of sentences in each paragraph. Then we compute the mean value of all paragraphs as our index and name it `CoREF`.

Deep Cohesion: This dimension reflects the degree to which a text contains causal and intentional connectives. To find the incidence of causal connectives, we first create a list of causal markers in the text. In particular, using a set of intra- and inter-sentence causal cues (Luo et al., 2016), we manually generate a list of regular expression patterns and use these patterns to find the causal connectives in a document. Then we compute the total number of causal connectives to the number of sentences in the document as a deep cohesion score. We call this index `PCDC`.

Verb Overlap: Based on FTT, abstract rather than concrete verb overlap across a text might help readers construct gist situation models. Wolfe et al. (2019) use two indices from Coh-Metrix to measure the verb overlaps in text including *SMCAUSlsa* and *SMCAUSwn*. Inspired by Coh-Metrix, we make some changes to improve these indices further. In particular, instead of LSA vectors, we leverage contextualized pretrained language models (PLMs) to get token vector embeddings to later compute the cosine similarity among them. Our

[3] https://github.com/UKPLab/sentence-transformers

[4] Model is available on Huggingface's hub by the name: `sentence-transformers/all-mpnet-base-v2`

hypothesis is that since PLMs have contextual knowledge of words in a text encoded in them, they may be a better choice than LSA for computing the vector representation of verbs in the text. We use spaCy's[5] transformer-based pipeline and the `en_core_web_trf` model which is based on *roberta-base* (Y. Liu et al., 2019) to compute token vector embeddings. Different forms of this index in GisPy follow the name pattern `SMCAUSe_*` where *e* stands for language model *embedding*.

To compute the WordNet verb overlap, we first find all synonym sets of verbs in a document in WordNet (Miller, 1995) with Part-of-speech (POS) tag *VERB*. Then for every pair of verbs, we check whether they belong to the same synonym set in WordNet or not. If yes, we assign score 1 to the verb pair, 0, otherwise. Then we compute the average of *1s* to the total number of sentences. Different implementations of this index follow the name pattern `SMCAUSwn_*`.

Word Concreteness and Imageability: To compute word concreteness and imageability (`PCCNC` and `WRDIMGc` in Coh-Metrix,) we use two different resources; (1) MRC Psycholinguistic Database Version 2 (Wilson, 1988), a resource that is used by Coh-Metrix and (2) word concreteness and imageability prediction scores using a supervised method introduced by Ljubešić, Fišer, and Peti-Stantić (2018).[6] For a document, we search document tokens based on their POS tags in these resources. Then we compute the average concreteness and imageability scores of all tokens in the document as the final scores. This process results in four scores in total named: `PCCNC_mrc, WRDIMGc_mrc, PCCNC_megahr, WRDIMGc_megahr` (two scores for each resource).

Hypernymy Nouns & Verbs: This index shows the specificity of a word in a hierarchy. The idea is that words with more hierarchy levels are less likely to help readers form gist inference than words with fewer levels (Wolfe et al., 2019). To compute this index, we first list all nouns and verbs in a document. Then for each word in the list, we find all synonym sets in the WordNet with the same POS tag (noun or verb). Then, we compute the average

[5] https://spacy.io/

[6] https://github.com/clarinsi/megahr-crossling

hypernym path length of all synonym sets of a word. The reason we find all synonym sets of a word instead of only one is that every word can have more than one synonym sets with the same POS and there is no way to know which synonym set has the same meaning as the word in the document. It would be fascinating to examine how we can identify the collection of synonyms that are most similar to a term in context as our future study.

4.1.3 Computing Gist Inference Score (GIS)

After computing all indices and before computing GIS which is a linear combination of these indices, we normalize all indices by converting them to z-scores because they can be on different scales. Then using the formula shown in Figure 4.2, we compute the final GIS for every document.[7] Documents with scores greater than zero in the positive direction have higher and smaller scores than zero in the negative direction have lower levels of gist, respectively.

4.2 Experiments

To test whether GisPy can correctly group and measure the level of gist in documents, we run our tool on a collection of datasets with known gist levels –low or high. We selected three benchmarks including two introduced by Wolfe et al. (2019) and one introduced by Broniatowski et al. (2016) to test the quality of scores in our tool. We gave more detail about these benchmarks in the following subsections. Before running GisPy, we also run Coh-Metrix on each dataset and compute GIS using the original Coh-Metrix indices. Our goal for doing so was to: (1) make sure we have a reliable gold standard that we can compare GisPy scores with and (2) reproduce the results from (Wolfe et al., 2019). Once we computed the GIS score using GisPy, to compare No Gist (low gist) vs. Gist (high gist) groups, we compared the mean of their GIS scores. Additionally, we conducted a

[7]We have implemented GisPy in a way that for exploring various ways of computing GIS using different linear combinations of variables, we can assign different weights to these variables.

t-test under the null hypothesis that there is no difference in the degree of gist between the two groups. The goal of running the t-test is to see whether our scores can significantly distinguish groups with lower and higher levels of gist.

Also, since for the five indices including Referential Cohesion, Verb Overlap based on Embeddings, Verb Overlap using WordNet, Concreteness, and Imageability we have multiple implementations, we computed the final GIS based on all possible combinations of these indices (320 sets of indices for each benchmark). Our goal was to find out what implementation of each index contributes better to distinguishing low vs. high gist documents. In a separate analysis, we also ran two robustness tests to ensure our results were not biased by seeing all possible combinations of indices.

4.2.1 Benchmarks

4.2.1.1 News Reports vs. Editorials

This benchmark includes 50 documents in two groups; (1) News Reports and (2) Editorials. Based on Wolfe et al. (2019), compared to News Editorials that provide a more coherent narrative, reports are more focused on facts. As a result, News Reports tend to have a lower level of gist than Editorials.

4.2.1.2 Journal Article Methods vs. Discussion

This benchmark includes 25 pairs of Methods and Discussion sections (total 50 text documents) from the same peer-reviewed scientific psychology journal articles. According to Wolfe et al. (2019), whereas the Methods part focuses on the interpretation of data, the Discussion section promotes replication of results. So, the Discussion portion ought to result in a higher gist score than the Methods section. This approach also controls for a number of variables such as author, journal, and topic.

4.2.1.3 Disneyland Measles Outbreak Data

Disneyland Measles Outbreak Data introduced by Broniatowski et al. (2016) also annotates gist. Documents in this dataset are articles (e.g., news) that are manually annotated by Amazon Mechanical Turk. There are a total of 191 articles with gist annotation among which there are *Gist-Yes: 147*, *Gist-No: 38*, and *unsure: 6* gist labels. We left out the *unsure* labels. Since full text of articles in this dataset were not available and each article only had a URL associated with it, we retrieved the full texts using the provided URLs. For those URLs that were no longer available, we used Wayback Machine to find the most recent image of the URL. At the end, we manually cleaned all articles and fixed the paragraph boundaries.

4.3 Results and discussion

Results of running GisPy on three benchmarks are shown in Tables 4.2, 4.3, and 4.4. For each benchmark, we list the top 10 combinations that most significantly distinguish low vs. high gist documents. As can be observed, when comparing indices with paragraph-level vs. non-paragraph-level implementations, paragraph-level indices typically produce better results. We do not necessarily observe a strong difference between local vs. global implementations. Additionally, we nearly always observe superior performance for concreteness and imageability indices when using *megahr* scores by Ljubešić et al. (2018). We leveraged *megahr* as a replacement for MRC that was originally used by Coh-Metrix.

Comparisons of individual indices for low vs. high gist documents from the best combination on each benchmark are shown in Figures 4.3, 4.4, and 4.5.

In Table 4.5, we also report the comparison of our best results on each benchmark with two baselines including (1) GIS computed based on the indices of Coh-Metrix, and (2) GIS reported by Wolfe et al. (2019). For the Disney benchmark, since we are the first work to report a baseline on this dataset, there are no other baselines. As can be seen, on the Reports vs. Editorials and Methods vs. Discussion we achieved performance on par with

Indices Combination									
PCREF	SMCAUSe	SMCAUSwn	PCCNC	WRDIMGc	**Low Gist**	**High Gist**	**Distance**	**t-statistic**	**p-value**
ap	1p	a	megahr	megahr	-3.842	-1.292	2.551	3.643	$* 7 \times 10^{-4}$
1p	1p	a	megahr	megahr	-3.833	-1.365	2.467	3.535	$* 9 \times 10^{-4}$
ap	1p	a	megahr	mrc	-3.850	-1.567	2.283	3.265	$* 2 \times 10^{-3}$
1p	1p	a	megahr	mrc	-3.840	-1.640	2.200	3.152	$* 3 \times 10^{-3}$
ap	1	a	megahr	megahr	-3.018	-0.830	2.189	3.216	$* 2 \times 10^{-3}$
ap	1p	a	mrc	megahr	-3.817	-1.662	2.155	2.967	$* 5 \times 10^{-3}$
1p	1	a	megahr	megahr	-3.009	-0.903	2.106	3.088	$* 3 \times 10^{-3}$
1p	1p	a	mrc	megahr	-3.807	-1.736	2.072	2.857	$* 6 \times 10^{-3}$
ap	1	a	megahr	mrc	-3.026	-1.104	1.921	2.792	$* 8 \times 10^{-3}$
ap	1p	a	mrc	mrc	-3.824	-1.937	1.887	2.429	$* 2 \times 10^{-2}$

$^{*}p \leq 0.05$

Table 4.2: Top 10 GIS scores computed for **Reports** (*Low Gist*) vs. **Editorials** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document.

Indices Combination									
PCREF	SMCAUSe	SMCAUSwn	PCCNC	WRDIMGc	**Low Gist**	**High Gist**	**Distance**	**t-statistic**	**p-value**
ap	1p	1	megahr	megahr	-0.282	4.730	5.012	7.188	$* 4 \times 10^{-9}$
ap	ap	1	megahr	megahr	-0.576	4.414	4.991	6.528	$* 4 \times 10^{-8}$
ap	1p	1p	megahr	megahr	-0.180	4.701	4.881	7.829	$* 4 \times 10^{-10}$
a	1p	1	megahr	megahr	-1.203	3.678	4.881	7.424	$* 2 \times 10^{-9}$
ap	ap	1p	megahr	megahr	-0.474	4.386	4.860	6.883	$* 10^{-8}$
a	ap	1	megahr	megahr	-1.497	3.362	4.860	6.460	$* 5 \times 10^{-8}$
1p	1p	1	megahr	megahr	-0.159	4.670	4.829	6.989	$* 8 \times 10^{-9}$
1p	ap	1	megahr	megahr	-0.453	4.355	4.808	6.328	$* 8 \times 10^{-8}$
a	1p	1p	megahr	megahr	-1.101	3.649	4.750	7.820	$* 4 \times 10^{-10}$
a	ap	1p	megahr	megahr	-1.395	3.333	4.729	6.594	$* 3 \times 10^{-8}$

$^{*}p \leq 0.05$

Table 4.3: Top 10 GIS scores computed for **Methods** (*Low Gist*) vs. **Discussion** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document.

and slightly better than Coh-Metrix. We achieved a significantly better distinguishment of low vs. high gist documents than what was reported by Wolfe et al. (2019). On Disney, GisPy significantly outperformed Coh-Metrix. These findings demonstrate that we not only replicated GIS indices, but in contrast to Coh-Metrix, we improved these indices in a fully open and transparent way. We hope this implementation transparency helps further improvement of these indices.

Indices Combination					Gist=No	Gist=Yes	Distance	t-statistic	p-value
PCREF	SMCAUSe	SMCAUSwn	PCCNC	WRDIMGc					
ap	1p	a	megahr	megahr	-1.921	0.497	2.418	3.440	$* 7 \times 10^{-4}$
1p	1p	a	megahr	megahr	-1.911	0.494	2.405	3.410	$* 8 \times 10^{-4}$
ap	1p	1	megahr	megahr	-1.729	0.447	2.176	3.158	$* 2 \times 10^{-3}$
1p	1p	1	megahr	megahr	-1.719	0.444	2.164	3.131	$* 2 \times 10^{-3}$
ap	1p	a	mrc	megahr	-1.676	0.433	2.109	3.416	$* 8 \times 10^{-4}$
1p	1p	a	mrc	megahr	-1.666	0.431	2.097	3.384	$* 9 \times 10^{-4}$
ap	1p	a	megahr	mrc	-1.652	0.427	2.079	3.415	$* 8 \times 10^{-4}$
1p	1p	a	megahr	mrc	-1.642	0.424	2.066	3.381	$* 9 \times 10^{-4}$
ap	1p	ap	megahr	megahr	-1.557	0.403	1.960	3.001	$* 3 \times 10^{-3}$
ap	ap	a	megahr	megahr	-1.554	0.402	1.956	2.950	$* 4 \times 10^{-3}$

$^{*}p \leq 0.05$

Table 4.4: Top 10 GIS scores computed by GisPy for **Gist=No** (*Low Gist*) vs. **Gist=Yes** (*High Gist*) articles in the Disney dataset. **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document.

Benchmark	Approach	Low Gist	High Gist	Distance	t-statistic	p-value
Reports vs. Editorials	GisPy	-3.842	-1.292	**2.551**	3.643	$* 7 \times 10^{-4}$
	Coh-Metrix	-4.148	-1.613	2.535	3.826	$* 3 \times 10^{-4}$
	(Wolfe et al., 2019)	-0.620	-0.252	0.368	-	-
Methods vs. Discussion	GisPy	-0.282	4.730	**5.012**	7.188	$* 3 \times 10^{-9}$
	Coh-Metrix	-2.077	2.933	5.010	6.331	$* 7 \times 10^{-8}$
	(Wolfe et al., 2019)	-0.297	0.45	0.747	-	-
Disney	GisPy	-1.921	0.497	**2.418**	3.440	$* 7 \times 10^{-4}$
	Coh-Metrix	-1.148	-0.151	0.998	1.878	6×10^{-2}

$^{*}p \leq 0.05$

Table 4.5: Comparison of GIS scores generated by GisPy vs. other baselines for all benchmarks.

4.3.1 Testing robustness

We conducted two tests to see whether our results are robust and generalize across all benchmarks.

Test 1: First, out of all combinations of indices, we separated those that significantly distinguished low and high gist groups in each benchmark resulting in 38, 281, 110 combinations for Report vs. Editorials, Methods vs. Discussion, and Disney benchmarks, respectively. We noticed that *all* combinations that are statistically significant in terms of t-test in Reports vs. Editorials benchmark are also statistically significant in the other two benchmarks. In

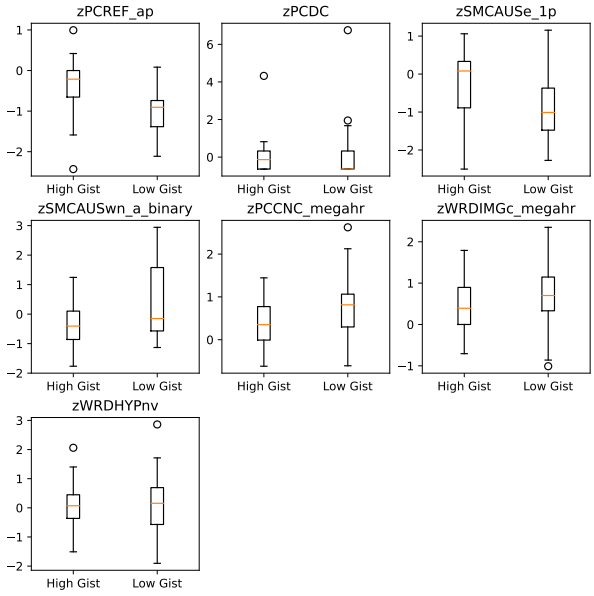


Figure 4.3: Individual indices of best GIS on *Reports (Low Gist) vs. Editorials (High Gist)*. All values are z-scores.

other words, there are 38 different combinations of indices that significantly distinguish low and high gist documents in *all* benchmarks. This confirms the robustness of indices implementation and their generalization.

Test 2: Second, we ran an extra experiment to ensure our best GIS scores on each benchmark are also robust when we do not know all possible combinations of indices to pick the best one. In particular, for each benchmark, using three different random seeds, we randomly split texts into a train and a test set each with balanced number of low and high gist documents. Then we computed GIS for documents in the train set and chose the best combination of indices that achieved the largest GIS distance between low and high gist groups. Then using that combination we computed GIS for documents in the test set. Results are reported in Tables 4.6. As can be seen, in all three benchmarks, the best indices combination on the train set also significantly distinguished the low and high gist documents in the test set. This further confirms that our GisPy indices are also robust when tested on unseen documents.

We also analyzed the individual indices from best combinations on the train set in robustness Test 2. These combinations are listed in Table 4.7. We noticed that for *PCREF*

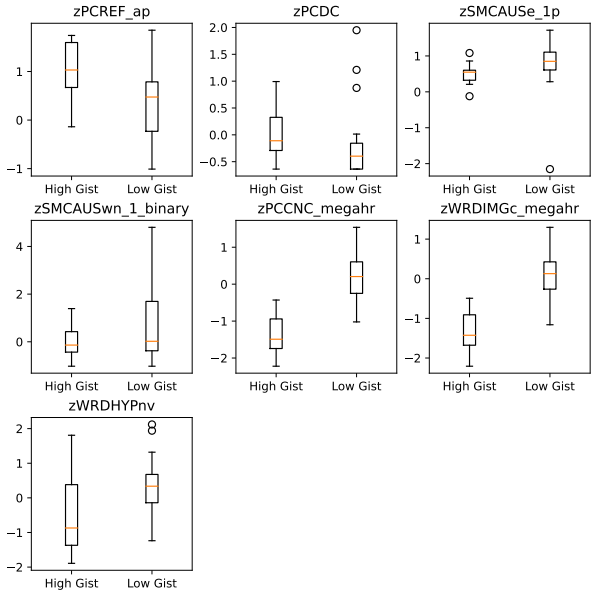


Figure 4.4: Individual indices of best GIS on *Methods (Low Gist) vs. Discussion (High Gist)*. All values are z-scores.

Benchmark	Train			Test		
	Low Gist	High Gist	p-value	Low Gist	High Gist	p-value
Reports vs. Editorials	-3.770	-1.131	$* 2 \times 10^{-2}$	-3.663	-1.413	$* 5 \times 10^{-2}$
Methods vs. Discussion	-0.634	4.538	$* 7 \times 10^{-5}$	-0.342	4.346	$* 2 \times 10^{-4}$
Disney Gist=Yes vs. Gist=No	-1.926	0.496	$* 4 \times 10^{-2}$	-1.910	0.493	$* 4 \times 10^{-2}$

$^{*} p \leq 0.05$

Table 4.6: GisPy GIS scores for train and test sets on all benchmarks.

and *SMCAUSe*, in 83% of the experiments, zPCREF_ap and zSMCAUSe_1p are part of the best combination. Also, for these two indices, in 89% of the times we obtained a better result using paragraph-level implementations than when we ignore paragraph boundaries. In other words, we obtain a better result by computing referential cohesion and semantic verb overlap using word embeddings at paragraph-level most of the time. For *PCCNC* and *WRDIMGc*, in all experiments with the exception of only one case only for *WRDIMGc*, scores computed by *megahr* achieved the best performance. And finally for *SMCAUSwn*, in 67% of the experiments, the global (*_*a*) implementation resulted in a better distinguishment between low and high gist documents than the local (*_*1*) implementation. Also, in only two experiments, the paragraph-level *SMCAUSwn* worked better than its non-paragraph-

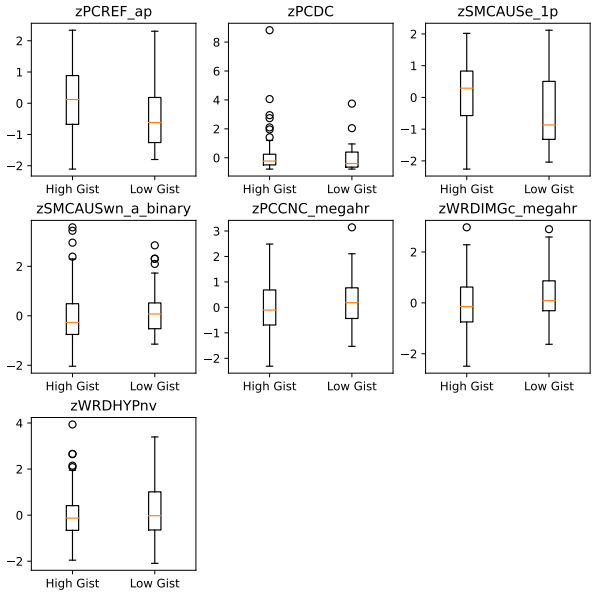


Figure 4.5: Individual indices of best GIS on *Disney Gist=No (Low Gist) vs. Gist=Yes (High Gist)*. All values are z-scores.

level implementation. Additionally, we dug a little deeper to understand why there is a difference between local vs. global *SMCAUSwn* across benchmarks. We noticed that the local indices only perform better in the Methods vs. Discussion dataset. We took a closer look to understand why this is the case. Interestingly, when we computed the ratio of the number of sentences to the number of paragraphs for all benchmarks, we observed that ratios for Reports vs. Editorials and Disney benchmarks, where global indices achieve a better performance, are **1.89** and **2.04**, respectively. And for Methods vs. Discussion where local indices perform better, the ratio is **6.48** which is significantly greater than the other two benchmarks. This may suggest that the density of paragraphs in terms of the number of sentences in each paragraph is one factor we need to keep in mind when selecting what implementation we want to choose for a benchmark. It would be interesting to run this analysis on more documents to see how our observation generalizes across different datasets.

Despite achieving significant improvements and solid results from robustness tests on three benchmarks from two domains, there is still great room to improve the quality of GisPy indices further. In this section, we will analyse challenges in the current implementation

Benchmark (S/P Ratio)	PCREF	SMCAUSe	SMCAUSwn	PCCNC	WRDIMGc
Reports vs. Editorials (1.89)	ap	1	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	mrc
	ap	1p	a	megahr	megahr
Methods vs. Discussion (6.48)	ap	1p	1	megahr	megahr
	ap	ap	1	megahr	megahr
	ap	1p	1	megahr	megahr
	a	1p	1p	megahr	megahr
	ap	ap	1p	megahr	megahr
	ap	1p	1	megahr	megahr
Disney (2.04)	ap	1p	a	megahr	megahr
	1p	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	1p	1p	a	megahr	megahr

Table 4.7: Best combinations in robustness Test 2 on the train set for all experiments separated by benchmark. S/P Ratio: the number of sentences to the number of paragraphs.

of GisPy and explain what we think can be a proper next step and direction in addressing them. We hope these insights inspire the community to keep working on this exciting line of research.

We did our best to bring three different benchmarks for measuring gist inference scores to life by aggregating, standardizing, and making them very easy to use. However, since measuring gist is a relatively newer and less investigated topic compared to readability, coherence, or cohesion, there is still a need for having higher quality benchmarks from different domains. The benchmarks we have tested our tool with are mainly from the news and scientific text domains. It would be interesting to see how our tool can be tuned on not only more documents from these domains but also other genres of text.

Also, our PCDC index, even though based on strong causal connective markers, mainly covers the explicit causal relations while not all causal relations are expressed explicitly in text. It would be interesting to think how we can enhance the quality of this index by

also including implicit relations and disambiguating causal connectives that can also be non-causal (e.g., temporal markers such as *since* or *after*) or leveraging discourse parsers such as DiscoPy (Knaebel, 2021).

We initially hypothesized that utilizing coreference resolution chains (`CoREF` index) may help us improve the referential cohesion index. By looking at the most significant combinations of indices in each benchmark, we noticed that `CoREF` appeared in 0/38, 53/281, 1/110 combinations for Report vs. Editorials, Methods vs. Discussion, and Disney benchmarks, respectively. As a follow-up, it would be interesting to see how coreference resolution can be leveraged in a different way —individually or in combination with other implementations of referential cohesion— to further improve this index.

4.4 Conclusion and future work

In this work, we introduced GisPy, a new open-source tool for measuring GIS in text. Evaluation of GisPy and robustness tests on three different benchmarks of low and high gist documents demonstrate that our tool can significantly distinguish documents with different levels of gist. We hope making GisPy publicly available inspires the research community to improve further indices of measuring gist inference in text.

Chapter 5: A Framework for Cause-Effect Relation Classification

Previous studies have shown the efficacy of knowledge augmentation methods in pretrained language models (PLMs). However, these methods behave differently across domains and downstream tasks. In this chapter, we investigate the augmentation of PLMs with knowledge graph data in the cause–effect relation classification and commonsense causal reasoning tasks. After automatically verbalizing triples in ATOMIC$^{20}_{20}$ and GLUCOSE, a wide coverage commonsense reasoning knowledge graph and a causal explanation dataset, respectively, we continually pretrained BERT and RoBERTa using Masked Language Modeling (MLM) objectives and evaluated the resulting models based on cause–effect pair classification and answering commonsense causal reasoning questions. Our results show that a continually trained pretrained language model (PLM) augmented with commonsense reasoning knowledge outperforms our baselines on two commonsense causal reasoning benchmarks, COPA and BCOPA-CE, and Temporal and Causal Reasoning (TCR) and PDTB3 datasets without additional improvement in model architecture or using quality-enhanced data for fine-tuning.

5.1 Introduction

The automatic identification of causal relations in text has been an important yet challenging task in Natural Language Understanding (NLU). Early methods in the 1980s and 1990s (Joskowicz et al., 1989; Kaplan & Berry-Rogghe, 1991; Garcia et al., 1997; Khoo et al., 1998) mainly relied on defining hand-crafted rules to find cause–effect relations. Starting in 2000, machine learning tools were utilized in building causal relation extraction models (Girju, 2003; Chang & Choi, 2004, 2006; Blanco et al., 2008; Do et al., 2011; Hashimoto et al., 2012; Hidey & McKeown, 2016b). Word embeddings and PLMs have also been leveraged in training models to understand causality in language in recent years (Dunietz et

al., 2018; Pennington et al., 2014; Dasgupta et al., 2018; Gao et al., 2019).

Investigating the true capability of PLMs in understanding causality in the text is still an open question. Recently, Knowledge Graphs (KGs) have been used in combination with PLMs to address commonsense reasoning. Two examples are CausalBERT (Z. Li et al., 2020), which is used for the guided generation of *cause* and *effect* and the model introduced by Guan et al. (2020) for commonsense story generation. Motivated by the success of continual pretraining of PLMs for downstream tasks (Gururangan et al., 2020), we explored the impact of commonsense knowledge injection into PLMs as a form of continual pretraining for causal reasoning and *cause–effect* relation classification. It is worth highlighting that even though researchers proved the efficacy of knowledge injection with continual pretraining for commonsense reasoning (Guan et al., 2020), the performance of these techniques is very dependent on the domain and downstream tasks (Gururangan et al., 2020). Moreover, to our knowledge, there are few studies on the effect of commonsense knowledge injection with knowledge graph data on cause–effect relation classification (Dalal et al., 2021). Our contributions are as follows:

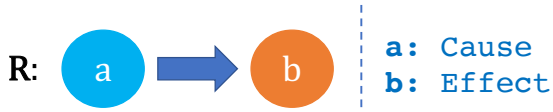
- We studied the performance of PLMs augmented with knowledge graph data in the less investigated cause–effect relation classification task.
- We demonstrated that a MLM framework using automatically verbalized knowledge graph data, without any further model improvement (e.g., new architecture or loss function) or quality-enhanced data for fine-tuning, can significantly boost the performance of cause–effect pair classification.
- We publicly released our knowledge graph verbalization codes and continually pretrained models.

The rest of this chapter is organized as follows: we start by defining some keywords and terminologies, including the definition of causal relations in Section 5.2. Then in Section 5.3, we review the history of automatic extraction of causal relations in text in detail

and different periods since the '80s, and we highlight some of the most well-known works in each period. Then we explain our method and framework in Section 5.4 followed by Section 5.5, where we list the benchmarks for evaluating our framework. We delineate the conducted experiments in Section 5.6 and report our results with an in-depth error analysis in Section 5.7. We conclude this chapter in Section 5.8.[1]

5.2 Definition of causal relations

Causal relations are interpreted differently, and there may not be a strong consensus on what they mean (Girju et al., 2002). Even though some of these interpretations are conceptually related, they may have fundamental differences. In this section, we explain what we mean by a *causal* or *cause-effect* relation in text or natural language.[2] Broadly speaking, by a causal relation, we mean a relation in which one state or event causes another state or event to happen or exist. To formulate this definition, by a causal relation we mean a relation, ***R***, between two entities, ***a*** and ***b***, where we call ***a*** and ***b*** two *elements*, *components*, or *arguments* of a causal relation. We demonstrate a causal relation like the following:



a causes b to happen/exist

where *R* is a causal relation in which *a* **causes** *b* to happen or exist. As a result, *a* is called the *cause*, and *b* is called the *effect*. There are a variety of features/properties associated with a causal relation. We explore these features in detail in Section 5.5 in our review of the datasets and annotation schemes of causal relations.

[1]This chapter contains a discussion of the work published in (Hosseini, Broniatowski, & Diab, 2022)
[2]We used the terms causal and cause–effect relations interchangeably in this thesis.

5.2.1 Causal relation categorization

There are different categories of cause–effect relations in text. These categories are defined in various ways in the literature. In this section, we start by reviewing these categories, followed by explaining what our definitions of these categories are.

- **Marked vs. Unmarked:** When there is a word in the sentence that signals the existence of the causal relation, we call it a *marked* causal relation. Otherwise, we call the relation *unmarked*.

- **Explicit vs. Implicit:** There is still an argument about when we should call a causal relation explicit and when we should call it implicit. In some work, such as that of Dasgupta et al. (2018), an explicit causal relation is defined as when both cause and effect arguments exist in the sentence versus an implicit relation in which either the cause or effect does not explicitly appear in the sentence or text. However, in some other work, implicit causal relations are defined as a causal relation when a reader needs to infer the relation even if both cause and effect exist in the text or sentence. As Mirza et al. (2014) also mentioned, *kill* is an example of such a definition because there is an embedded effect which is *death* in it. According to Pitler and Nenkova (2009), implicit relations are the relationships inferred to exist by a reader in the absence of a discourse connective. They mention that identifying implicit relations is much harder than identifying explicit relations. Even though such a definition is not necessarily for causal relations, as far as causal relations are a sub-type of relations in text, it would be fair to apply such a definition to them as well.

- **Our definition:** In this thesis, by explicit causal relations, we mean that cause and effect arguments are both present in the form of a token or multiple tokens in the document while for an implicit relation, one of the arguments does not exist in the document. Furthermore, if there is any word or connective to signal the causal relation, then we call the relation marked. Otherwise, we call it unmarked.

It is also worth noting that arguments of a causal relation can be defined in various ways. These arguments could be a single token/word or multiple tokens/words in the language. From a grammatical point of view, these arguments can be *noun phrases (NPs)*, *verb phrases (VPs)*, *clauses*, *events*, *named entities*, or even *full sentences*. Specifying the type of arguments, *a* and *b*, in a relation is challenging and still an open problem in causal relation extraction.

5.2.2 Causal relation identification tasks

The identification of causal relations can be categorized into two general groups:

- **Causal Relation Extraction:** Given a text document as input, the output of a causal relation extraction task is a collection of pairs of *(cause, effect)*, where *cause* and *effect* are a sequence of tokens in the document, and there is a causal relation between them. Figure 5.1 shows an example of causal relation extraction.

- **Causal Relation Classification:** In causal relation classification, we assume arguments of a relation are already detected or extracted. Moreover, our task is to specify whether there is a causal relation between these arguments. Alternatively, knowing there is a causal relation between the arguments, the task can be defined as detecting the cause or effect (specifying the *direction*) in the relation.

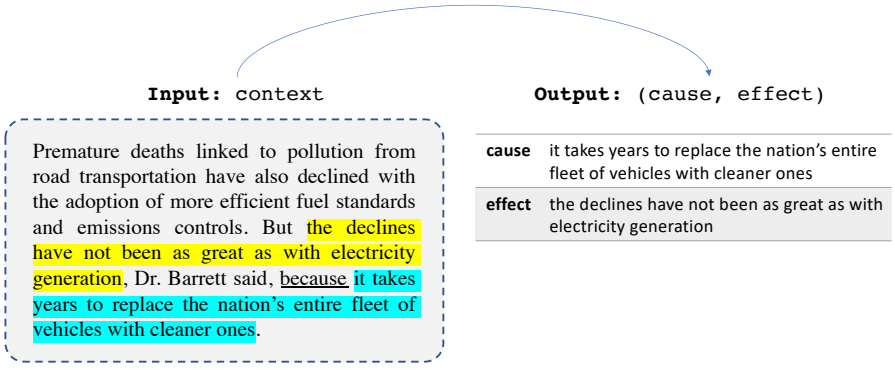


Figure 5.1: Example of input and output in causal relation extraction.

5.3 History of automatic causal relation identification in text

Efforts to automatically detect causal relations in text started a couple of decades ago. To better demonstrate how research around this topic evolved, we review works about cause–effect relations from a historical point of view and in chronological order in different periods. In particular, we break this section into three parts: (1) Before 2000, most works on causal relation identification relied on hand-coding and rule-based methods. (2) From 2000 to 2018, machine learning-based solutions started to appear in developing systems for identifying causal relations in combination with rule-based approaches. Also, since 2013, in line with the bigger trend in NLP research, deep learning began to play a role in developing different methods for cause–effect relations identification. Another fact in this period was that compared to works before 2000, research was conducted in a highly principled manner concerning the evaluation of the performance of methods and models. (3) From 2018, the year that contextualized [pretrained] language models such as BERT (Devlin et al., 2018) were introduced and started a new trend in NLP and Natural Language Understanding (NLU), until now.

Covering all works in the past couple of decades around causal relation identification can be the topic of a separate study by itself. There is always a chance that we miss some work. However, here, we try our best to cover the works that represent the general theme of research on identifying causal relations in each period.

5.3.1 Before 2000: hand-coding and rule-based methods

The automatic extraction of causal relations[3] in text roughly started in the late '80s and early '90s. One of the earliest works on automatic identification of causal relations was introduced by Joskowicz et al. (1989), later mentioned by Khoo (1995) in their dissertation as one of the two only computer programs for causal relation extraction along with Selfridge,

[3]Studying causal relations in the text may have been started before the 80s, but here we mainly focus on automatic extraction of these relations using computational methods and tools.

Daniell, and Simmons (1985). In their work, Joskowicz and colleagues introduced a system named PROTEUS to extract the causal relations between facts in short messages in the Causalty Reports (CASREPs) about equipment in Navy ships. One of the main insights from this work was the importance of domain knowledge in not only causality extraction but in language processing. The design of PROTEUS heavily relied on the specific case of a simulation model of equipment in ships. This work also presented the idea of hypothesis generation for facts that are dangling and are not part of any causal chain. In such cases, they mentioned, the simulation model would not be enough for finding causal relations.

Two years later, Kaplan and Berry-Rogghe (1991) introduced a causal analyzer that was part of TAKT, a tool for acquiring knowledge from text. Their analyzer, which in contrast to PROTEUS, was domain-independent, aimed to construct a sequence of causally related events. The architecture of TAKT is shown in Figure 5.2. Their approach has three causal recognition processes: a recognizer for explicit and implicit causal instances, a model-based recognizer for causality, and a constraint-based recognizer. One point here is that based on Kaplan and Berry-Rogghe (1991)'s definition, explicit causal relations are not necessarily relations that are marked in the text. By explicit, they likely meant that arguments of a causal relation explicitly exist in the sentence. Another interesting part of Kaplan and Berry-Rogghe (1991)'s work was constructing a causal chain using pairs of causal events in a sentence. Typical of papers from that era, there is no statistical analysis of the model's performance, and their method is only on two hand-coded texts.

The mid- to late-1990s was when we saw some more principled efforts at text-based causality extraction using automatic, yet simple, methods. Khoo and colleagues (Khoo, 1995) are part of this era. Extracting cause–effect pairs expressed in sentences using linguistic patterns is one of the research questions in Khoo's dissertation (Khoo, 1995). In their work, they mainly used linguistic markers and patterns without domain knowledge to extract *explicit* cause–effect pairs in a collection of Wall Street Journal news articles (Khoo et al., 1998). They did not use knowledge bases because they wanted to develop a method

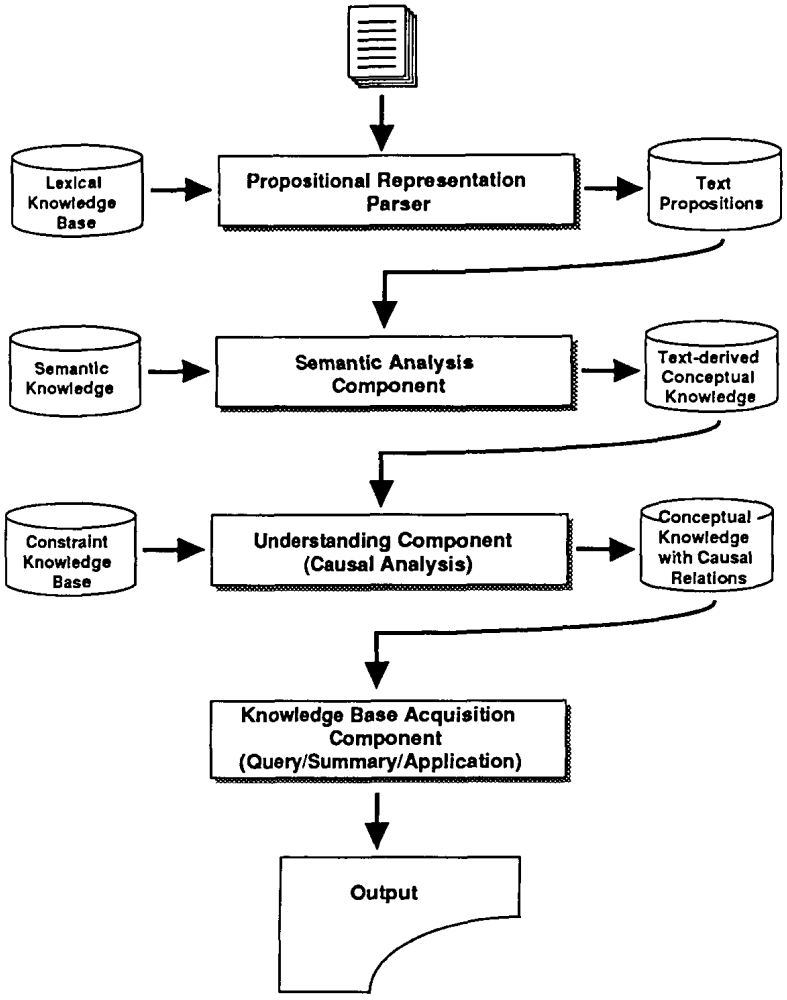


Figure 5.2: Architecture of TAKT. Image from (Kaplan & Berry-Rogghe, 1991).

independent of external resources that can generalize well and be used in other subject matters. Another argument for them for not using the knowledge bases was that, as they say, back in the time, there were only a few available knowledge bases for certain domains, so they did not want to develop a method which is limited by the lack of external resources. In their work, they identified five ways of explicitly expressing cause–effect pairs: (1) using causal links, (2) using causative verbs, (3) using resultative constructions, (4) using conditional constructions, and (5) using causative adverbs and adjectives where the fifth is not used in their work for what is mentioned as the lack of studies and a comprehensive list of such adverbs and adjectives. The scope of text in their work for extracting causal relations

is within a sentence and adjacent sentences. They did a pretty interesting error analysis with some helpful insights. One insight is that they say that one reason for false-positive samples, or as they say, commission error, is sense ambiguity. They summarize the main sources of errors in their system as (1) complex sentence structure, a problem that can be solved using better parsers, (2) lexical ambiguity, and (3) absence of inferencing from world knowledge.

There are some other efforts in the 90s to automatically locate causal relations in language, including the work by Garcia et al. (1997). In their work, Garcia extracted the relations between *actions* as expressed by verbs in French, and their approach is mainly based on using verbs as indicators of causality in text. Twenty-three relations of causality, including efficient causality and formal causality, were used in their system, COATIS. COATIS analyzed the text for causality extraction sentence by sentence and mainly worked based on a strategy called Contextual Exploration, which was triggered by indicators of causality. One of the main features of COATIS that Garcia highlighted was that COATIS is independent of any domain and can extract causality in any technical text without needing domain knowledge. It is also interesting that to this point in the history of causal relation extraction, there was no comparison with other works, and we can barely find any number or statistics regarding the performance of the proposed methods and algorithms. Garcia says their system did not make much noise and mentioned that quantifying what their system did not find, what we know as recall, is more difficult. Another interesting point about Garcia's paper was how they tried to go beyond theory and develop a system that extracts causal relations. Reconstruction of the original figure of the semantic organization of the relations of efficient causality (extract) in Garcia's paper is shown in Figure 5.3. Efforts on automatic causal relation, as Garcia also mentioned, were still at early stages in that period until 2000.

5.3.2 2000-2018: machine learning starts to play a role

In early 2000, Khoo et al. (2000) presented an improved version of their previous work. Their system still heavily relied on the causal indicators/signals and linguistics patterns

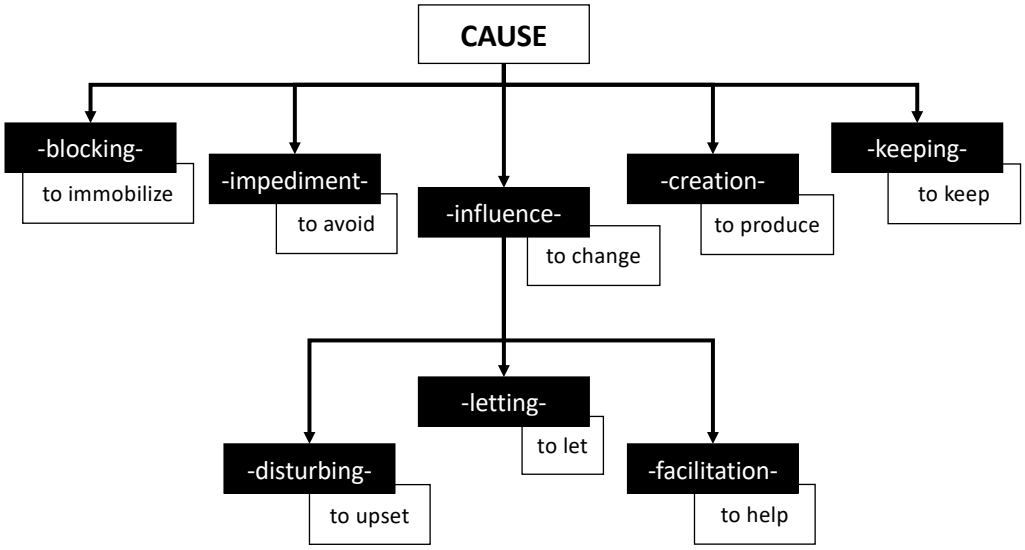


Figure 5.3: Reconstruction of the original figure of the semantic organization of the relations of efficient causality (extract) in Garcia's paper.

to extract explicit relations, this time from the Medline journal abstracts, but they used full parsing to get the syntactic structure of sentences instead of just using part-of-speech tagging and phrase bracketing as opposed to their previous work. They came up with some graphical patterns to find the causal relations. As they mentioned in their error analysis as well, a noticeable number of errors in their system was still coming from the fact that the syntactic parser they used was not good enough. Khoo and colleagues' studies are among the first work that tried to automatically extract causal relations in different domains in a more principled manner. Their relatively more rigorous error analysis gave the community helpful and interesting insights that remain relevant even after all these years.

A year later, Low et al. (2001) introduced a semantic expectation-based causation knowledge extraction methodology, SEKE. One assumption SEKE made was that causes and consequences mostly appear in the same sentence or close to each other. SEKE was studied to analyze the Hong Kong Stock Market to find reasons that affect the movements of the Hang Seng Index (HSI). Considering how SEKE was evaluated, it may be fair to reason that SEKE had the potential to over-fit. For example, most of the wrongly-classified samples in SEKE's evaluation were the result of unseen reason factors, suggesting that SEKE, even

on a specific domain, could not generalize well. SEKE can be described as what the authors also mentioned when they concluded their paper as "a *workable* approach." It was *workable* but had its limitations. In line with what Joskowicz et al. (1989) demonstrated, Low et al. (2001) showed that even when we humans try to extract relations such as causal relations in a domain, we need to have knowledge about that domain and know what we expect to read to be better equipped to identify causal relations.

In 2002, a new line of research on the identification of causal relations was started by Girju and colleagues (Girju et al., 2002). The main area of focus in their work was finding intra-sentential relations based on a syntactic pattern in the form of $< NP1\ verb\ NP2 >$ where the verb is a simple causative verb. This work's definition of explicit and implicit causal relations is somehow different than what we defined and discussed earlier. For example, in a sentence such as "John caused the disturbance," the causal relation was not considered explicit because John did not directly cause the disturbance, but his action did. Girju et al. (2002)'s method had two main steps, including (1) discovering lexico-syntactic patterns that can express causal relations and (2) validating and ranking the ambiguous patterns based on semantic constraints on nouns and verbs (e.g., using WordNet 1.7 (Miller, 1995) to come up with a list of verbs which are ranked based on their ambiguity and frequency). The results of this work were mainly compared to those of Garcia et al. (1997) and Khoo et al. (2000). The way Girju et al. (2002) compared their results with previous work did not seem to be highly relevant because they applied their method to a different dataset, and their definition of causal relations was not necessarily consistent with previous work. That way of comparing results to a certain degree was also representative of how different works, at least on causal relation identification, were being compared to one another in that period —not necessarily in a principled or rigorous manner. Regardless of the limitations of their work, Girju et al. (2002)'s approach to leveraging an ontology such as WordNet in finding linguistic patterns of causal expression was rather novel and interesting.

Not long after their previous work, Girju, for the first time, employed a learning-based

method to find lexico-syntactic patterns that demonstrate causality in text (Girju, 2003). The basic idea of this work was that not all the noun phrases in the $< NP1\ verb\ NP2 >$ are associated with causal links. The main difference between this and their previous work was using a learning-based method, the C4.5 algorithm, to train a model to find lexico-syntactic patterns. This work also showed why causal relations are important in Question Answering, one of the most important downstream tasks in NLP. This was when we saw more people trying to use learning-based methods to tackle the problem of causal relation extraction.

Chang and Choi also tried to extract causal relations between NPs using what they called cue phrases (Chang & Choi, 2004, 2006). They used the Naive Bayes classifier to classify ternary candidates in form of $< cause\ NP,\ cue\ phrase,\ effect\ NP >$. These ternary candidates were generated using the Connexor dependency parser (Tapanainen & Jarvinen, 1997) and ultimately were filtered using a list of cue phrases which were the causal verbs introduced in Girju and colleagues' work (Girju et al., 2002). One of Chang and Choi's contributions was using an unsupervised method to learn the probability of cue phrases and lexical pairs.

Later in 2008, Blanco and colleagues introduced their work on extracting marked and explicit causal relations (Blanco et al., 2008). Soon after reading their work, some similarities between this work and Girju's works (Girju et al., 2002; Girju, 2003) can be noticed, which is not a surprise when we realize both Blanco and Girju worked with Moldovan. Blanco's work also used a combination of rule-based and learning-based methods. They started by finding syntactic patterns (i.e., $[VP\ rel\ C]$, $[rel\ C,\ VP]$) and extracted the instances of such patterns in a collection of 1,270 sentences. Then they trained a machine learning model using a set of features they defined to classify the pattern instances as either causal or non-causal. One nuance and interesting point in Blanco and colleagues' work was that they were perhaps one of the first works which brought up a point about comparing results with other works and addressing the fact that comparison of results is not always easy. In particular, they mentioned that it was difficult to compare their work with others because they used a

specific pattern different from other works for causal relation extraction. Another insight from this work was the degree to which different causal connectors signal causal relations in text. For example, the fact that "because" signals causation in more than 98% of the time could help us rely on this connector in a distant supervision method to generate samples of causal relations. It is worth mentioning that in this work, only syntactic information was used in the causal tagging, and investigation of semantic features was mentioned as one next step to improve the introduced method.

Do et al. (2011) introduced a minimally supervised system for finding causal relations between the events in a sentence as well as the adjacent sentences in a defined window. The input to their system is in form of *p(a_1,...,a_n)* where *p* is a predicate that triggers an event and *a* is an argument of the event. Causality between events in this work was measured using a formula called the Cause-Effect Association (CEA) that had three components, including (1) association between event predicates, (2) association between the predicate of an event and the arguments of the other event, and (3) association between event arguments. Also, by leveraging resources such as WordNet (Miller, 1995) and FrameNet (Baker, Fillmore, & Lowe, 1998), authors created a list of the event triggering nouns and used Penn Discourse Treebank (PDTB) to identify discourse relations of types comparison, contingency, expansion, and temporal. As one of the most highly-cited works in this area, it is worth highlighting some insights from this work. First, it is interesting from a historical point of view in research around causal relation identification that this paper mentions that as of 2011, authors were not aware of any benchmark for evaluating event causality extraction. Hence, they created their evaluation benchmark. Second, one of the annotation guidelines in this work is that the *cause* event in a causal relation should temporally precede the *effect* event, an assumption that later was shown not to be necessarily the case all the time. Third, this work also highlighted the need for additional contextual knowledge because the lack of such knowledge was one main source of errors for both false positives and false negatives in their experiments. Lastly, one novelty of this work was combining event causality prediction

and discourse relation prediction through a global inference procedure and using Integer Linear Programming (ILP).

Starting from 2012, we also see a series of works by Hashimoto and colleagues on causality extraction in text (Hashimoto et al., 2012). The main idea of this work is *excitation* as a semantic orientation that facilitates the identification of causality and some semantic relations, including contradictions between pairs of events. Excitation has three categories: excitatory, inhibitory, and neutral. The basic idea was that in the real world, too, activation or deactivation of something can result in activation or deactivation of something else, and based on this, we may be able to use excitation to extract causality. In other words, their main assumption was that if there is a strong excitation tendency between two templates connected by AND/THUS-type connectives, one of these templates tends to have a causal effect on the other template. Once they built a network of excitation templates, their method extracted causality in pairs connected to non-causal and AND/THUS-type connectives. They used connectives that did not explicitly signal causal relations to evaluate how useful excitation templates were in detecting causality.

Two years later, in 2014, Hashimoto and colleagues also presented a system for predicting future scenarios based on event causality for Japanese (Hashimoto et al., 2014). They claimed their method can be applied to other languages as well and is not limited to only Japanese. This future hypothesis generation work was an improved version of their 2012 work and aligned with their previous efforts. One nuance in Hashimoto and colleagues' work was their definition of causality. They looked at causality probabilistically, not logically, meaning that if A causes B, A *increases the chance* for B to happen, not necessarily causes B to happen. Another interesting point they highlighted was verifying event causality pairs in the real world. One of the assumptions, for example, they took into consideration when annotating samples of event causality in their supervised system was that event causality pairs should be self-contained, meaning that the reader should be able to infer causality in a pair without any kind of contextual knowledge or information. The introduced method

in this work has two main steps, including (1) extracting event causality candidates and (2) defining a set of features and ranking the candidates. In the first step, a collection of excitation templates was used for extracting candidates that were pairs of events in sentences. The main idea of the causality extraction phase was based on the hypothesis that event causality involves some type of semantic relations between the phrases in a pair. Based on this hypothesis, they defined a fairly large collection of patterns for different semantic categories such as CAUSATION, MATERIAL, and PREVENTION and used them as their semantic features. They also defined a set of context- and association-based features—among which Do et al. (2011)'s CEA measure is used—and studied their effects on their model performance in an ablation study. One year later, Hashimoto et al. (2015) introduced another framework for generating event causality hypotheses. The primary motivation behind their work was that not all causal relations in the world are written in text corpora, no matter how large. They started by generating a set of hypothesis candidates in their method and ranking them based on plausibility using an SVM classifier. From a historical point of view, one interesting point about this paper was the idea of *generating* hypotheses, especially if we consider the fact that Natural Language Generation (NLG) was not as trendy as it is today in 2015. Recently, we see more works such as COMET(-ATOMIC) (Hwang et al., 2020) that also try to achieve the same capability in language models to see whether a model can predict or generate the likely causes and effects of world phenomenon.

In (Sorgente et al., 2013) a hybrid system was designed and implemented for extracting explicit causal relations from text. In the introduced approach, which was a combination of rule-based and learning-based methods, and in the first step, sentences that potentially contained pairs of cause and effect were extracted from text using regular expressions and a lexico-syntactic pattern matching approach. Then using rules defined for each of the patterns, the exact cause and effect pair were extracted. Then in the second part, a classifier was trained using lexical, semantic, and dependency features to filter extracted pairs that were not a causal relation.

Paramita Mirza and colleagues also have a line of research on causality extraction starting in 2014. One theme of their research was the interaction between causality and temporal relations. Their line of work roughly started in 2014 by adding causal annotations to the TempEval-3 corpus (Mirza et al., 2014). They presented an annotation scheme for tagging explicit causal relations between events in their work. Annotation of causal links was restricted to the presence of explicit causal constructions such as basic constructions (cause, enable, and prevent), expressions containing affect and link verbs, and periphrastic accusatives. As part of this annotation scheme, polarity, factuality, and certainty were also annotated. As part of this work, a system for automatically tagging explicit causal relations in the text was also presented to test the soundness of the annotation scheme. The SemEval 2013 (UzZaman et al., 2013) task data collection was used where events were annotated in documents. One interesting part of this study was investigating the overlap between causal and temporal relations. Their results showed that in almost one-third of the extracted causal relation, there is also a temporal relation. This overlap could have been higher had they also considered the implicit causal relations. The high ambiguity of some connectors in text, such as "since" for causal relations, was also investigated in this study. In the same year, Mirza and colleagues also introduced a system for extracting temporal and causal relations between events (Mirza, 2014) followed by another work to analyze the interactions between these relations (Mirza & Tonelli, 2014). In their work, two classifiers were trained, one for finding causal signals in text and another for detecting causal links/relations between pairs of events. Their method covered both inter- and intra-sentential causal links. Like their previous work, they also studied the effect of temporal relations in improving causal relation extraction. The authors also introduced a new dataset, Causal TimeBank, by adding causal links and causal signal information, CLINK and C-SIGNAL, respectively, to the well-known TimeBank dataset. Two years later, in 2016, Mirza and Tonelli introduced a new system of causal and temporal relation extraction from text—CATENA, (Mirza & Tonelli, 2016).

Luo et al. (2016) introduced a way to extract causal relations between short texts,

including phrases and sentences. This work outlined two directions in causality extraction, which were in some way representative of the trends until 2016 in the method for causality identification. One direction was based on extensive manual efforts in extracting causal relations, which resulted in high precision. It also included a deep lexical and syntactic analysis, which usually resulted in low coverage. Their goal was to have both high precision and a high recall. One interesting concept in their method was about *necessity* causality and *sufficiency* causality and using these two metrics to define a causality strength score. These two metrics were also two critical components in the formula mentioned in (Van den Broek, 1990) to measure the strength of a causal relation. Luo et al. (2016)'s work was also different from previous work in that it tried to look at causality from a commonsense perspective. This work also introduced a causal network, **CausalNet**, as a large graph-based representation of general causality knowledge that can provide the relatively reasonable computation of causal strength between terms.

Caselli and colleagues have also contributed to the research around causality extraction in text in a series of valuable works. Event StoryLine Corpus (ESC) project is one of these contributions. Development of ESC started from a proposal on a storyline annotation and representation scheme (StaR) by Caselli and Vossen (2016) for a new task they defined as StoryLine Extraction (Caselli & Vossen, 2017). StaR was, in fact, an extension of ECB+ (Cybulska & Vossen, 2014) and was built for grounding a group of narratological concepts to linguistic elements in documents. Relations in StaR, including causal and temporal relations, were between the events. The goal of their task was to extract events spread over time and relevant to stories. As one of the main components in their dataset, they defined *ACTION_CAUSATIVE* for annotating events describing causal relations. For events in their annotation, they kept the head of a phrase unless for a few exceptions, e.g., when an event had multiple words such as *American Civil War*. Another link in their corpus related to causal relations was *PLOT_LINKs*. As described, PLOT_LINKs were related to both causal and additional event–event relations, such as contingency or entailment, and

they aimed at answering "why" something happened. As parts of PLOT_LINKs annotation, *rising_action* and *falling_action* were used to label cause and consequences. In their work, they annotated both explicit and implicit relations at intra- and inter-sentential levels.

As part of this period, distant supervision in causal relation extraction also started to be employed as a popular method, especially considering the fact that generating gold standard data for this task is not easy. Hidey and McKeown (2016a)'s work on introducing a distant supervision-based method for identifying causality is one example. Also, at the end of this period, we started to see new problems and tasks being defined around causal reasoning and causality extraction in language. In (Kang et al., 2017) for example, a new task called causal explanation generation was proposed where the idea was that given a target (effect), y, and a source (cause), x, how can we generate a time series of events that explain the causal relationship between the source and target. This task also has some similarities with Hashimoto et al. (2014)'s work, as also mentioned by the authors.

5.3.3 2018-Present: the dominance of deep neural networks

2018 was the year we started to see more works and methods based on deep neural networks. Dasgupta et al. (2018) implemented an automatic method using Bi-LSTM neural models to extract intra-sentence level causal relations from text. One interesting feature of this work was annotating and finding the cause and effect relations in a sentence separately and at the word level, unlike other works that mainly find causal relations as a binary task or cause and effect as a single word. In particular, each word in an input sentence was annotated by E (effect), C (cause), causal connective (CC), or N (none) tags. One of the contributions of this work was also creating a new dataset of causal relations by combining some existing data resources but a new annotation schema. From the historical point of view, it is also interesting to see that Dasgupta et al. (2018) still complained about the lack of annotated resources for causal relation extraction. This work shows that even though deep learning had already been used in NLP since around 2013, in causal relation extraction, it still was a

new approach.

Jointly studying temporal and causal relations was another interesting line of research that continued to be worked on in this period. In another work by Ning et al. (2018a) these two relations were studied jointly, and it was shown that learning temporality and causality together could improve the performance of each of these two tasks as well.

More recently, there were also some attempts for causal relation extraction among all the events at the document level. Gao et al. (2019), for example, presented a model of the causal relations within and across the sentences in a document. There was no direction prediction in the models. They just classified a pair of event mentions as either causal or not. Another example of a method for extracting intra-sentence causal relations is the work by Phu and Nguyen (2021). The intuition behind the design of their model is building interaction graphs among useful information and objects in a document and then using these interaction graphs in Graph Convolutional Networks (GCN).

Even though, after 2018, we mainly observed the dominance of deep learning methods, rule-based models for extracting causal relations were not wholly abandoned. Eidos (Sharp et al., 2019) was an instance of such systems for extracting causal relations from free text. Eidos worked based on finding the causal signaling words in the text. In particular, in its first step, using the signal words and Odin's grammar, a set of initial cause and effect concepts were found. Then using the dependency syntax of the input document, these initial cause and effect concepts were expanded.

In (P. Li & Mao, 2019), prior human knowledge was used in a Convolutional Neural Network (CNN) model called K-CNN to create a system for detecting cause–effect relations. The model is composed of two channels, including knowledge-oriented and data-oriented channels, respectively. Data-oriented channel learns from data while the knowledge-oriented channel generates better linguistic clues and keywords using external resources, WordNet, and FrameNet, to augment prior human knowledge to help the model better find causal relations, which is interesting. In K-CNN, one of the assumptions is that two events are

Paper	Method	Relation	Pair	Data
(Joskowicz et al., 1989)	rule-based	both	yes	no
(Kaplan & Berry-Rogghe, 1991)	rule-based	both	no	no
(Garcia et al., 1997)	rule-based	explicit	no	no
(Khoo et al., 1998, 2000)	rule-based	explicit	yes	no
(Low et al., 2001)	rule-based	explicit	yes	no
(Girju et al., 2002; Girju, 2003)	combined	explicit	yes	no
(Chang & Choi, 2004, 2006)	combined	explicit	yes	no
(Blanco et al., 2008)	learning-based	explicit	no	no
(Do et al., 2011)	combined	both	yes	yes
(Sorgente et al., 2013)	combined	explicit	yes	no
(Mirza et al., 2014)	rule-based	explicit	no	no
(Mirza & Tonelli, 2014)	learning-based	explicit	no	yes
(Sharp et al., 2019)	rule-based	both	yes	no

Table 5.1: Summary of some of the works on causal relation identification in natural language.

already annotated in a sentence, and we want to figure out whether there is a causal relation between these events or not.

A summary of some of the works we covered in this section and their comparison is shown in Table 5.1. The *method* column in this table specifies what method (e.g., rule-based, learning-based, or combined) is used in the introduced approach. Also, the *relation* column demonstrates the type of causal relation identified by the method (e.g., explicit, implicit, or both). The *pair* column contains a binary value of yes or no to indicate whether the exact cause/effect statements/text spans are extracted by the method. Lastly, the "Data" column shows whether the work introduced a new dataset for the causal relation extraction task. It is also worth pointing out that some of the recent advancements in identifying causal relations in natural language, including efforts based on leveraging Transformer-based PLMs are covered in detail by Drury et al. (2022) and Yang et al. (2022).

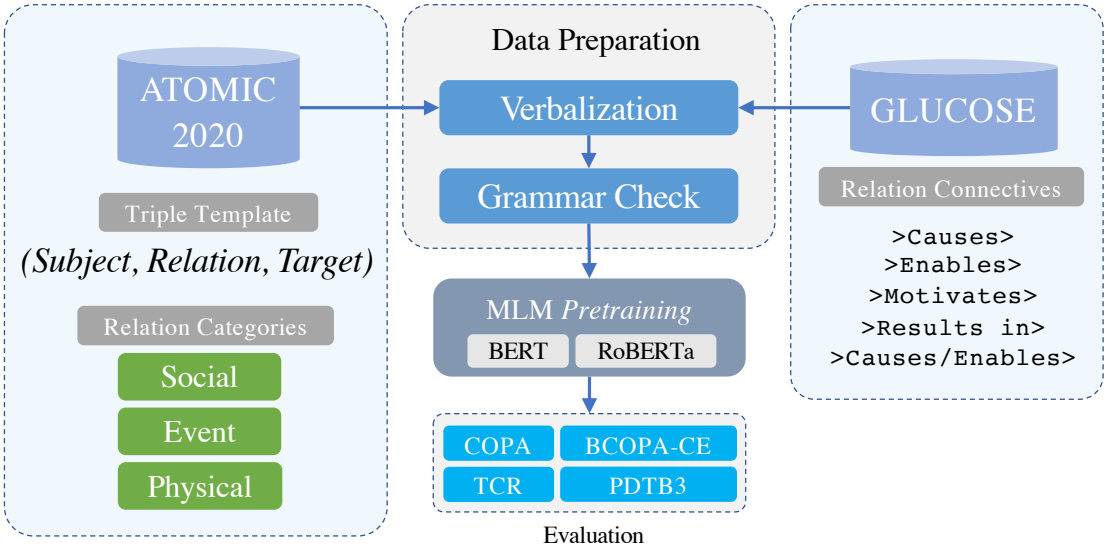


Figure 5.4: Overview of our proposed framework to continually pretrain PLMs with commonsense reasoning knowledge.

5.4 Method

The overview of our method is shown in Figure 5.4.[4] Our method contains three main steps: (1) Preparing input data using KGs for continual pretraining, (2) Continual pretraining, and (3) Evaluating the resulting models on cause–effect relation classification benchmarks. In the data preparation step, we converted data from two well-known commonsense and causal reasoning KGs, including GLUCOSE (Mostafazadeh et al., 2020) and ATOMIC$^{20}_{20}$ (Hwang et al., 2021), to natural language text.[5] Figure 5.5 shows the distribution of sequence length of data in GLUCOSE and ATOMIC$^{20}_{20}$. After verbalizing the KGs data, we continually pretrained BERT (Devlin et al., 2018) and RoBERTa (Y. Liu et al., 2019) using MLM, and in the end, evaluated the performance of the resulting models using different benchmarks. In the following, we explain each of these steps in further detail.

[4]Codes and models are publicly available at: https://github.com/phosseini/causal-reasoning.
[5]In the literature, such a conversion process is called *verbalization* as well (Agarwal et al., 2021b)

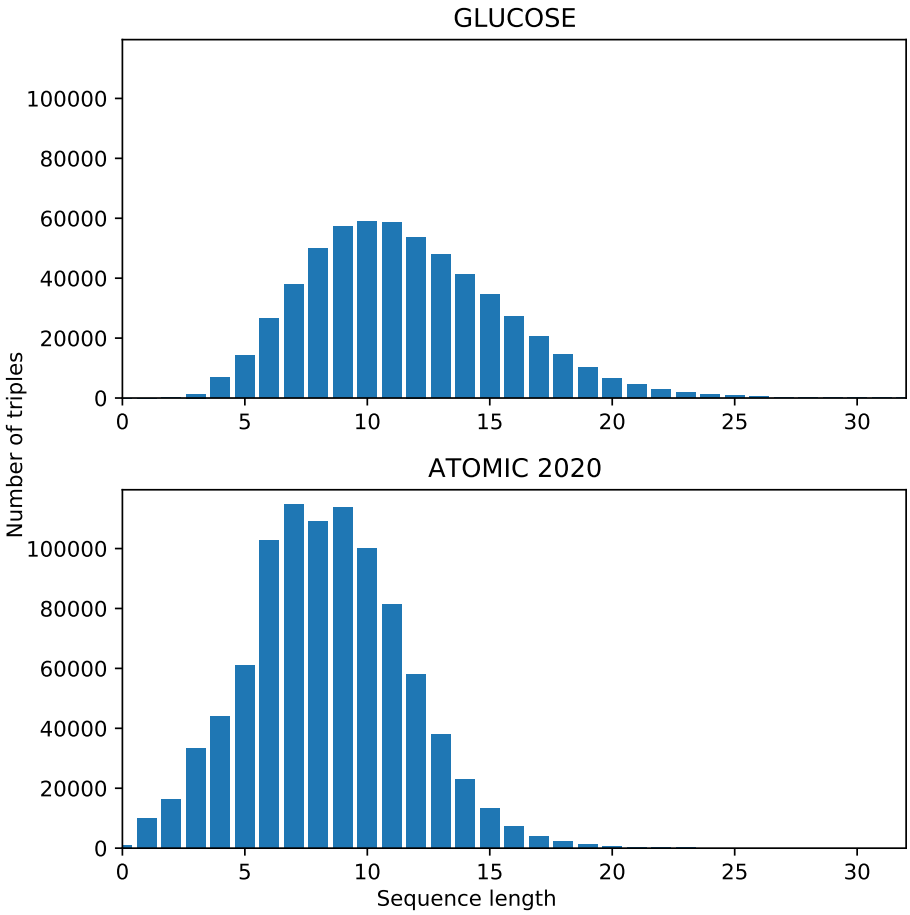


Figure 5.5: Distribution of samples in GLUCOSE and ATOMIC$^{20}_{20}$ based on the number of tokens (separated by space).

5.4.1 ATOMIC$^{20}_{20}$ to text

Samples in ATOMIC$^{20}_{20}$ are stored as triples in the form of *(head/subject, relation, tail/target)* in three splits, including train, development, and test. ATOMIC$^{20}_{20}$ has 23 relation types classified into three categorical types, including commonsense relations of social interactions, physical-entity commonsense relations, and event-centric commonsense relations. In the rest of this chapter, we refer to these three categories as social, physical, and event, respectively.

Filtering Triples: We removed all duplicates and ignore all triples in which the target value is none. Moreover, we ignore all triples that include a blank. Because we need to know the gold value of the masked tokens in MLM, a triple that already has a blank (masked

token/word) may not help our pretraining. For instance, in the triple: `[PersonX affords another ___, xAttr, useful]` it is hard to know why or understand what it means for a person to be useful without knowing what they afforded. This preprocessing step yields 782,848 triples with 121,681, 177,706, and 483,461 from the event, physical, and social categories, respectively. The distribution of these relations is shown in Figure 5.6.

Converting Triples: Each relation in ATOMIC$^{20}_{20}$ is associated with a human-readable template. For example, *xEffect*'s and *HasPrerequisite*'s templates are *as a result, PersonX will*, and *to do this, one requires*. We use these templates to convert triples in ATOMIC$^{20}_{20}$ to sentences in natural language by concatenating the subject, relation template, and target. Examples of converting triples to text are shown in Figure 5.7.

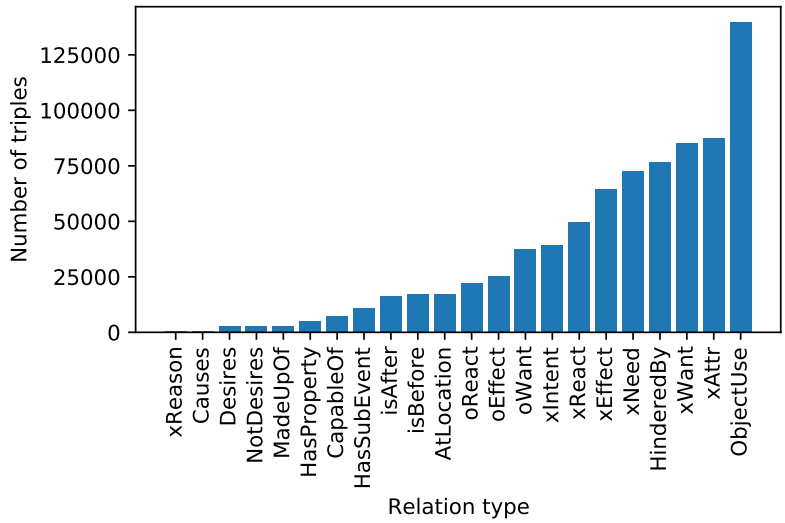


Figure 5.6: Distribution of relation types in ATOMIC$^{20}_{20}$.

Checking Grammar: When we convert triples to natural language text, ideally, we want to have grammatically correct sentences. Human-readable templates provided by ATOMIC$^{20}_{20}$ are not necessarily rendered in a way to form error-free sentences when concatenated with subject and target in a triple. To address this issue, we use an open-source grammar and spell checker, LanguageTool,[6] to double-check our converted triples to ensure they do not contain obvious grammatical mistakes or spelling errors. Similar approaches that include

[6] https://tinyurl.com/yc77k3fb

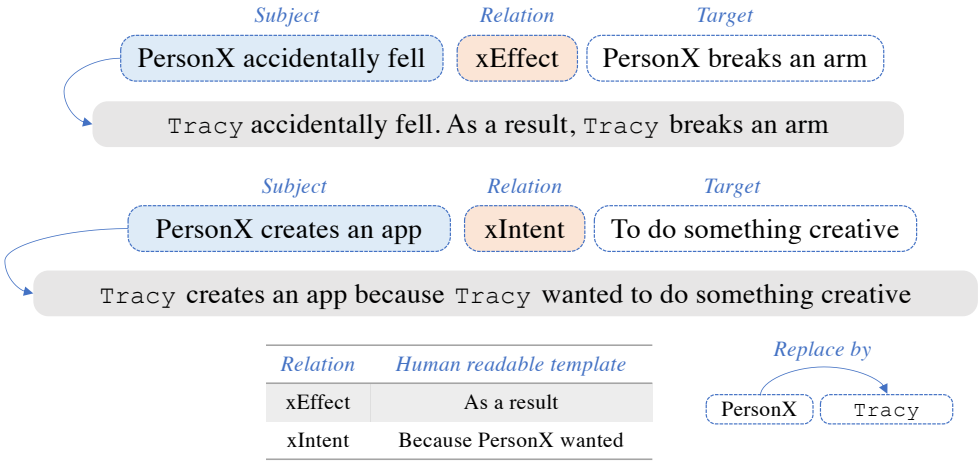


Figure 5.7: Examples of converting two triples in ATOMIC$^{20}_{20}$ to natural language text using human-readable templates. Following Sap et al. (2019), we replace *PersonX* with a name.

deterministic grammatical transformations were also previously used to convert Knowledge Graph (KG) triples to coherent sentences (Davison et al., 2019). It is worth pointing out that data-to-text generation (KG verbalization) is a separate task, and efforts have been made to address this task (Agarwal et al., 2021a). We leave investigating the effects of using other data-to-text and grammar-checking methods to future research.

5.4.2 GLUCOSE to text

GLUCOSE is a large-scale dataset of implicit commonsense causal knowledge. Each data point in GLUCOSE includes 10 dimensions of causal explanation focusing on events, states, motivations, and emotions. Out of these 10 dimensions, half are specific causal statements, and the remaining half are general rules that capture the implicit commonsense knowledge. An example of a data point in GLUCOSE is shown in Table 5.2.

Using a slightly modified version of templates that are provided for causal connectives (CCs) in GLUCOSE, we concatenate the two spans in a causal relation with each relation's template to form a verbalized sample. The CCs in GLUCOSE include: `[>Causes/Enables>, >Motivates>, >Enables>, >Causes>, >Results in>]`.

Story	**Sadie was on the volleyball team**. She had the ball in her hands. She went to spike it over the net. Suddenly, the ball smacked another girl in the face. Sadie apologized but was asked to leave the court.
Dimension	**Specific Statement/Template**
specificNL	Sadie tries out for the volleyball team >Causes/Enables> Sadie is on the volleyball team
specificStructured	Sadie_[subject] tries out_[verb] for _[preposition1] the volleyball team_[object1] >Causes/Enables> Sadie_[subject] is_[verb] on _[preposition1] the volleyball team_[object1]
generalNL	Someone_A tries out for Something_A (that is team) >Causes/Enables> Someone_A is part of Something_A
generalStructured	Someone_A_[subject] tries out _[verb] for _[preposition1] Something_A ‖that is team‖_[object1] >Causes/Enables> Someone_A_[subject] is part_[verb] of _[preposition1] Something_A_[object1]

Table 5.2: Example of an entry in GLUCOSE

5.4.3 Continual pretraining

As mentioned earlier, we used MLM to continually pretrain our models. We followed the default procedure in BERT and RoBERTa[7] to create the input data for our pretraining (e.g., number of tokens to mask in input examples). For ATOMIC$^{20}_{20}$, we run the pretraining using ATOMIC$^{20}_{20}$'s *train* and *development* functions were used to split our training and evaluation sets, respectively. For GLUCOSE, because there is only a training set, we split the training set into *training* (90%) and *development* (10%) sets. For both datasets, we continually train our model for 15 epochs by default on Google Colab TPU v2 using the *PyTorch/XLA* package with a maximum sequence length of 30[8] and batch size of 128. We evaluated our models every 500 steps on the evaluation set, and to avoid overfitting, we used early stopping with the patience of 5 on evaluation loss. Ultimately, we selected the best model as the one with the least evaluation loss. For clarity, we added *ATOMIC-* and *GLUCOSE-* prefixes to the name of models that are trained with ATOMIC$^{20}_{20}$ and GLUCOSE data, respectively.

[7]We use Huggingface's *BertForMaskedLM* and *RobertaForMaskedLM* implementations.
[8]99.99% of ATOMIC$^{20}_{20}$ and GLUCOSE instances have 30 tokens or fewer.

5.5 Benchmarks

When finding causal relation benchmarks and datasets, we noticed that existing benchmarks, even though valuable, are somewhat scattered. Because these scattered resources do not necessarily follow the same format and schema, we decided to aggregate and unify them by introducing CREST—a schema for cause–effect relations in text. In the rest of this section, we first review some of the most well-known causal relation datasets and benchmarks. Then we explain and introduce our schema CREST and its features in more detail.

Two of the well-known and widely-used resources of causal relations were introduced in the SemEval 2007 task 4 (Girju et al., 2007a), and SemEval 2010 task 8 (Hendrickx et al., 2009). These two datasets have a total of 1,442 samples of causal relations. Events in these relations are tagged with binary labels indicating whether there is a causal relation between the two annotated events in the relation or not. In another work, Bethard et al. (2008) introduced a corpus of 1,000 samples of causal and temporal event pairs annotated on top of some of the Wall Street Journal articles in the PDTB. Event pairs in this corpus are limited to those connected using *and* in documents. Figure 5.8 shows one example of annotation in this corpus.

```
<link id="811">
<label type="causal">CAUSAL</label>
<label type="temporal">BEFORE</label>
<annotation file="wsj_0426" leaf="2" sentence="0" />
<annotation file="wsj_0426" leaf="7" sentence="0" />
</link>
```

Figure 5.8: A sample from (Bethard et al., 2008) that shows the annotated causal and temporal labels of an event pair.

As one of the early attempts to tag causal relations in text, there are a couple of interesting insights in this work. One of these insights is about exploring temporal features to see if verb tenses can aid in the detection of causal relations. It suggests that designing a theory about how temporal and causal relations interact may be helpful. A similar idea is also the subject of the study in (Mirza et al., 2014) where Mirza and colleagues analyzed CLINKs and

TLINKs in their work and demonstrated that from the total number of CLINKs that they had, around 33% of them have underlying TLINKs. Richer Event Description (RED) (O'Gorman et al., 2016) annotation scheme and corpus is another example of a resource that annotates multiple semantic relations, including causal and temporal relations as well. Another insight from Bethard et al. (2008)'s work was that syntactic and lexical information may not be enough for extracting causal relations, and semantic information is also necessary for this task.

PDTB (Prasad et al., 2008) is another well-known corpus with causal relation tags. Causal relations are among the semantic relations under the category of *contingency* sense. One of the features of PDTB is having the sense tag for connectives which makes PDTB an ideal resource for relation sense ambiguation tasks as well. In PDTB, a sense tag is attached to each relation to clarify what type of relation a connector indicates (e.g., specifying whether *since* is a causal or temporal connector in a relation).

Mostafazadeh et al. (2016) also introduced a scheme called CaTeRS for annotating temporal and causal relations in text. CaTeRS mainly focuses on those temporal and causal relations that support learning stereotypical narrative structure, and it is not intended to be a comprehensive analysis of all temporal and causal aspects of the text. The introduced framework in CaTeRS captures a set of event entities in stories and their pairwise semantic relations. Similar to Mirza et al. (2014), CaTeRS also uses the type of causal relations, *Cause*, *Enable*, *Prevent* defined by Wolff (Wolff, 2007).

Caselli and Vossen (2017) introduced a new benchmark called ESC for temporal and causal relation detection. ESC contains pairs of events and relations between these events. And events are defined based on the same definition as in (Mostafazadeh et al., 2016; Mirza et al., 2014). In addition to the causal relation detection task, ESC is potentially an accurate benchmark for testing systems and generating a coherent message and mental representation of documents, including news as well. ESC is designed to be used as a benchmark in three tasks, including event detection and classification, the temporal anchoring of events, and

explanatory relation (e.g., causal) identification and classification. Temporal relations in this corpus are marked using TLINKs, and explanatory relations such as causal relations are marked using PLOT_LINKs. The total number of identified explicit causal relations in is 117. Moreover, the data format in ESC is based on CAT, a tool for content annotation.

BECauSE Corpus 2.0 is one of the most recent corpora of causal relations (Dunietz et al., 2017a). BECauSE 2.0 contains explicit causal relations and is an improved version of BECauSE 1.0 (Dunietz et al., 2015). This corpus contains a total number of 1803 annotated sentences as causal samples. One of the main features and improvements of the latest version of BECauSE is extending causal language annotation to adjacent semantic relations instead of annotating samples with only one type of semantic relation as in BECauSE 1.0. BECauSE is one of the few corpora and annotation schemes that mainly focuses on tagging causal relations. However, it has some room for improvement, such as handling edge cases and tagging *implicit* causal relations.

5.5.1 CREST: A causal relation schema for text

Causal relation benchmarks and datasets are scattered and do not often follow the same schematic or machine-readable format. Such incompatibility among resources is not a new phenomenon, and it has slowed progress in the classification of semantic relations, including causal relations (Hendrickx, Kim, Kozareva, Nakov, Séaghdha, et al., 2010). That is why we decided to unify a collection of widely-used resources of causal relations, including those we covered at the beginning of this section, to make them accessible for use in the research community by introducing CREST.

In this step, we focused on publicly available datasets in which the *context* associated with a causal relation is available, and we excluded data resources in which a relation is only a pair of words/tokens without providing any context such as (Luo et al., 2016). The main reason is that causal relation/pair classification is mainly a context-dependent task. We also excluded data resources containing only context and a label (binary or multi-label)

without annotating the spans/arguments of a relation such as (Yu, Li, & Wang, 2019). It is worth pointing out that these excluded resources can fit in the CREST's format, although they were just not our focus for now. It is also important to mention that some resources are not publicly available, which is itself another challenge in addition to the incompatibility of schemes, such as Prasad et al. (2008), and some others, such as the corpus of temporal-causal relations (Bethard et al., 2008) and BECauSE (Dunietz et al., 2017b) only share their annotation.[9]

CREST is a schema or, more precisely, a unifying framework for storing causal/non-causal relations. Various formats and schemes are introduced for annotating causal relations (Bethard et al., 2008; Prasad et al., 2008; O'Gorman et al., 2016; Mostafazadeh et al., 2016). There are two main reasons for the existence of such a wide range of schematics: (1) some data resources, (Mirza et al., 2014) for instance, are created on top of existing datasets built for tasks besides causal relation extraction, so they have to follow the same format as the original dataset, and (2), some recent datasets include newly-defined meta-information for relations that were not annotated in older datasets. For example, SemEval-2007 task 4 (Girju et al., 2007b) and SemEval-2010 task 8 (Hendrickx, Kim, Kozareva, Nakov, Séaghdha, et al., 2010) data do not have annotations for causal signal or trigger words. However, such words are shown to be effective in causal relation extraction (Do et al., 2011) and were later annotated in other datasets such as BECauSE (Dunietz et al., 2017b).

By comparing a wide range of well-known and widely-used datasets of causal relations, we identified the largest common set of features and annotations among them to define CREST. In Table 5.3, we listed all these features. We call the process of converting relations in a dataset to CREST as *CRESTing* the dataset.

[9] E.g., not all the annotated samples in (Bethard et al., 2008) and (Dunietz et al., 2017b) can be used unless we have access to Wall Street Journal and The New York Times Annotated corpora.

(Field name: Field value type)	Explanation
`global_id: string`	Identifier of a relation in CREST.
`original_id: string`	Identifier of a relation in the original dataset.
`span1, span2, signal: list`	These three fields store the arguments of a relation, cause, effect, span (for a non-causal relation), and signal/trigger tokens. The value of these fields is a list of strings.
`context: string`	The context in which a relation appears.
`idx: string`	Character indices of *span1*, *span2*, and *signal* tokens in the *context*. Each index is stored in a separate line in the following format: *span_name*[space]*start_index*:*end_index*. E.g., *span1 14:29*
`label: integer`	Specifies whether there is a causal relation between *span1* and *span2*. 0: non-causal. 1: causal
`direction: integer`	Shows the direction of a relation. 0: *span1* $\Longrightarrow$ *span2* and 1: *span1* $\Longleftarrow$ *span2*
`split: integer`	Split to which a relation belongs in the source dataset. (0: train, 1: dev, or 2: test set).

Table 5.3: List of features in CREST. All *value types* are Python values.

5.6 Experiments

We chose multiple benchmarks in our experiments to ensure we thoroughly tested the effects of our newly trained models. These benchmarks include (1) the TCR dataset (Ning et al., 2018b), which is a benchmark for the joint reasoning of temporal and causal relations; (2) the COPA (Roemmele et al., 2011) dataset, which is a widely-used and notable benchmark (Rogers et al., 2021) for commonsense causal reasoning; (3) BCOPA-CE (Han & Wang, 2021), which is a new benchmark inspired by COPA that contains unbiased token distributions, which makes it a challenging benchmark to use; and (4) PDTB3. For COPA-related experiments, because COPA does not have a training set, we use COPA's development set for fine-tuning our models and testing them on COPA's test set (COPA-test) and BCOPA-CE. For hyperparameter tuning, we randomly split COPA's development set into a training set (90%) and a development set (10%) and found the best learning rate,

batch size, and the number of training epochs based on the evaluation accuracy of the development set. Then we fine-tune our models and evaluate them on the test set using COPA's original development set and best set of hyperparameters. For TCR, we respectively fine-tuned and evaluated our models on training and test splits.

Because the samples in the PDTB3 experiment were not separated into training, development, and test groups by default, we created these splits. To ensure we split these samples not just randomly but in a way that resulted in the minimum semantic overlap across samples in different splits, we measured the semantic overlap between samples after randomly splitting them into training, development, and test sets. We used `Sentence Transformers` (Reimers & Gurevych, 2019) to compute vector embeddings, and then we found the similarity between all pairs of samples in splits (e.g., train vs. dev. or train vs. test) by calculating the cosine similarity between the samples' vectors. Next, in every pair of splits, we filter out the highly similar sample from the larger set in the pair that has more than 50% similarity with any of the samples in the other set. We always filter out samples from the larger set to ensure we do not lose too many samples from dev and test sets when compared to the train set. This process results in 994, 394, and 467 samples for our train, dev, and test sets.

In the COPA-related experiments, we used *BERT-Large-Cased* and *RoBERTa-Large* pretrained models in all experiments as our baselines. We converted all instances to SWAG-formatted data (Zellers et al., 2018) and used Huggingface's *AutoModelForMultipleChoice* which is a model with a multiple-choice classification head on top. Also, when encoding the samples and before running our experiments on COPA, we ran an additional experiment to see if or how including the question type, `asks-for="cause"` or `asks-for="effect,"` can potentially be helpful in boosting the models' performance. The results of this experiment are shown in Table 5.4. As can be seen, adding question type as a prompt does not necessarily help the BERT-Large baseline model achieve better performance. Hence, for now, we decided not to add question type as a prompt. Two points are is worth noting

here: (1) this setting can also be viewed as a way of testing models in understating *implicit* causality because models will make predictions without hinting about the question and answer types, and (2) these results are better not interpreted as adding a prompt does not help. Prompt engineering is receiving much attention in the NLP research community, and it may be worth investigating better prompts for this task for future work.

Prompt for `"cause"`	**Prompt for** `"effect"`	**Acc (%)**
It is because	As a result	$69.9_{(0.07)}$
What was the cause of this?	What happened as a result?	$66.5_{(0.05)}$
No prompt	*No prompt*	$\mathbf{75.5}_{(0.07)}$

Table 5.4: Results of adding COPA question types as Prompt vs. No Prompt on BERT-Large baseline.

For TCR and PDTB3, we converted every instance by adding special tokens to input sequences as event boundaries and fine-tune and test the R-BERT [10] model (Wu & He, 2019). We chose R-BERT for our relation classification because it leverages the pretrained embeddings and transfers information about target entities (e.g., then events in a relation) through the model's architecture and incorporates encodings for the target entities. Also, because R-BERT's architecture is only compatible with the BERT configuration, for now, we run our TCR and PDTB3 experiments only with the BERT model. Examples of COPA, TCR, and PDTB3 are shown in Figure 5.9. BCOPA-CE has the same format as COPA.

To ensure our results are robust and consistent and are not biased by only picking the best performance, we run all experiments using eight random seeds and report the average accuracy and standard deviations. Even though we will list the performance of other work on the same benchmarks, the comparison of results may not be entirely fair because, in some studies, either only the highest performance is reported, or it is not clear if the reported results are reproducible and robust if we use different random seeds for initializing models and data shuffling.

[10]We use the following implementation of R-BERT: https://github.com/monologg/R-BERT. Also, our fork of R-BERT with minor modifications can be found at: https://github.com/phosseini/R-BERT

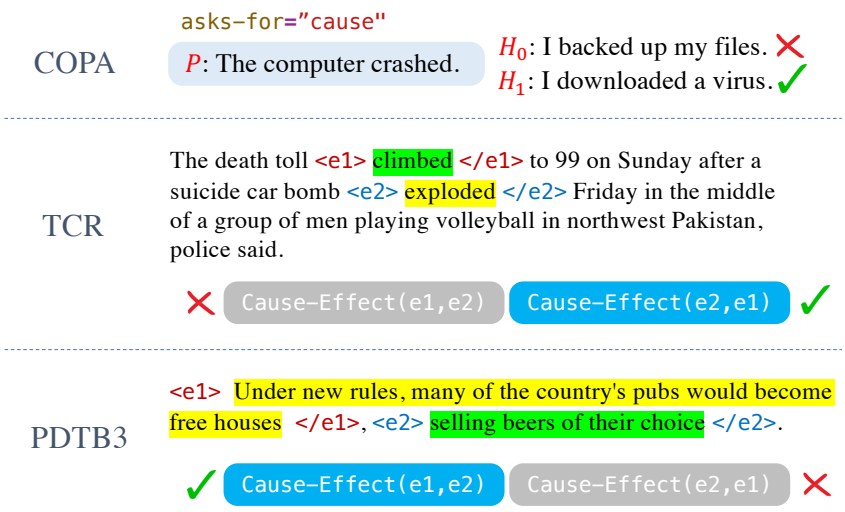


Figure 5.9: Examples of COPA, TCR, and PDTB3.

5.7 Results and discussion

Results of our experiments on TCR and PDTB3 are shown in Table 5.5 and Table 5.6. As can be seen, for TCR, our models significantly outperform our baseline. We also achieved a better performance than the joint inference framework by Ning et al. (2018b) who formulated TCR as an ILP problem. On PDTB3, although our ATOMIC-BERT model achieved performance on par with the baseline, our GLUCOSE-BERT model outperformed both the baseline and the ATOMIC-BERT models. As mentioned earlier, we tried to remove any semantic similarities between our training, dev, and test data in PDTB3, and we resolved the possible overlaps among these sets to ensure we do not have any leaks. Still, as for future work, it may be worth designing better splits from PDTB3 for the task of cause–effect relation classification to investigate further the impact of commonsense knowledge augmentation for causal reasoning on this dataset.

The results of the experiments in COPA-test are shown in Table 5.7. For ATOMIC-related experiments, we decided to train two separate models, one time using all three types of relations and another time only using the *event* relations. The reason is that event-centric relations in ATOMIC$^{20}_{20}$ contain commonsense knowledge about event interactions

Model	Acc (%)	Best Acc (%)
Joint system (Ning et al., 2018b)	77.3	-
BERT-Large (baseline)	$73.4_{(0.05)}$	80.0
ATOMIC-BERT-Large	$77.6_{(0.07)}$	86.0
GLUCOSE-BERT-Large	$\mathbf{80.2}_{(0.09)}$	91.0

Table 5.5: TCR Accuracy results.

Model	Acc (%)	Best Acc (%)
BERT-Large (baseline)	$80.6_{(0.03)}$	84.0
ATOMIC-BERT-Large	$80.5_{(0.11)}$	87.0
GLUCOSE-BERT-Large	$\mathbf{83.0}_{(0.03)}$	86.0

Table 5.6: Penn Discourse Treebank Version 3.0 (PDTB3) Accuracy results.

for understating likely causal relations between events in the world (Hwang et al., 2021). In addition, event relations have a relatively longer context (# of tokens) than the average of all three relation types combined, which means more context for a model to learn from. As a result, we wanted to see whether continually training a model with less but more relevant data with a longer context achieves a different performance. As can be seen, all our ATOMIC and GLUCOSE models significantly outperformed the baselines. Although the improvement for BERT models compared to the baseline is nearly 5%, the performance enhancement for RoBERTa models is even larger and more than 10%. These results demonstrate the positive effect of augmenting PLMs with commonsense knowledge for causal reasoning and identifying cause–effect relations. Also, *Event Only* ATOMIC models achieve performance on par with and, in the case of RoBERTa, even better than the models trained with all three categories of relations and more data. These results suggest that the relevance of data to our task for continual pretraining may also be an important factor when choosing our training data.

Three points are worth mentioning here. First, our models, BERT-Large and RoBERTa-Large, have a significantly lower number of parameters than state-of-the-art models, Google T5-11B (∼32x) and DeBERTa-1.5B (∼4x) and it shows how smaller models can be com-

Model	Acc (%)
PMI (Roemmele et al., 2011)	58.8
b-l-*reg* (Han & Wang, 2021)	71.1
Google T5-base (Raffel et al., 2019)	71.2
BERT-Large (Kavumba et al., 2019)	76.5
CausalBERT (Z. Li et al., 2020)	78.6
BERT-SocialIQA (Sap et al., 2019)*	80.1
Google T5-11B (Raffel et al., 2019)	94.8
DeBERTa-1.5B (He, Liu, Gao, & Chen, 2020)	96.8
Our Models	
BERT-Large (baseline)	$75.5_{(0.07)}$
ATOMIC-BERT-Large	
- Event, Physical, Social	$79.1_{(0.03)}$
- Event only	$79.1_{(0.01)}$
GLUCOSE-BERT-Large	$\mathbf{79.9}_{(0.02)}$
RoBERTa-Large (baseline)	$74.1_{(0.11)}$
ATOMIC-RoBERTa-Large	
- Event, Physical, Social	$83.9_{(0.02)}$
- Event only	$84.9_{(0.03)}$
GLUCOSE-RoBERTa-Large	$\mathbf{85.7}_{(0.03)}$

Table 5.7: COPA-test Accuracy results.

petitive and benefit from continual pretraining. Second, we have not yet applied any model improvement methods such as using a margin-based loss introduced by Z. Li et al. (2019) and used in CausalBERT (Z. Li et al., 2020), an extra regularization loss proposed by Han and Wang (2021), or fine-tuning with quality-enhanced training data, Balanced COPA (BCOPA), introduced by Kavumba et al. (2019). As a result, there is still great room to improve current models that can be a proper next step. Third, we achieved performance that was on par with BERT-SocialIQA (Sap et al., 2019), [11] and we did not use expensive crowdsourcing or any *manual* re-writing or correction to verbalize KG triples to create our pretraining data.

We ran an additional experiment on the *easy* and *hard* question splits in COPA-test that were categorized by Kavumba et al. (2019) to determine how our best models perform on

[11] Our best random seed run achieved 81.4% accuracy.

harder questions that do not contain superficial cues. The results are shown in Table 5.8. All our models significantly outperformed the baselines on both easy and difficult questions. The RoBERTa models achieved a larger performance improvement than their baseline compared to the BERT models.

Model	Easy	Hard
BERT-large-FT (Kavumba et al., 2019)	$83.9_{(0.04)}$	$71.9_{(0.03)}$
RoBERTa-large-FT (Kavumba et al., 2019)	$91.6_{(0.01)}$	$85.3_{(0.02)}$
Our Models		
BERT-Large (baseline)	$84.7_{(0.05)}$	$69.8_{(0.09)}$
ATOMIC-BERT-Large		
- Event, Physical, Social	$90.6_{(0.02)}$	$72.1_{(0.03)}$
- Event only	$88.6_{(0.02)}$	$73.2_{(0.02)}$
GLUCOSE-BERT-Large	$89.1_{(0.02)}$	$74.2_{(0.03)}$
RoBERTa-Large (baseline)	$80.5_{(0.11)}$	$70.2_{(0.12)}$
ATOMIC-RoBERTa-Large		
- Event, Physical, Social	$87.5_{(0.02)}$	$81.7_{(0.03)}$
- Event only	$90.7_{(0.03)}$	$81.3_{(0.04)}$
GLUCOSE-RoBERTa-Large	$89.6_{(0.05)}$	$83.3_{(0.03)}$

Table 5.8: COPA-test Accuracy results on Easy and Hard question subsets.

Lastly, the results of experiments on BCOPA-CE are shown in Table 5.9. As expected, based on the results also reported by Han and Wang (2021), we initially observed that our models are performing nearly as random baseline. Because we did not use the type of question when encoding input sequences, we decided to see whether doing so would improve the performance of the models. We added `"it is because"` and `"as a result"` as prompt to `asks-for="cause"` and `asks-for="effect,"` respectively. Interestingly, the new model outperforms the baseline and Han and Wang (2021)'s *b-l-aug* model that is fine-tuned with the same data as ours when question types are added as prompts to input sequences of correct and incorrect answers in the test set.

Model	Acc (%)
b-l-*aug* (Han & Wang, 2021)	51.1
b-l-*reg* (Han & Wang, 2021)	64.1
Our Models	
BERT-Large (baseline)	$51.5_{(0.01)}$
ATOMIC-BERT-Large	
- Event only	$53.2_{(0.01)}$
- Event, Physical, Social	$53.5_{(0.02)}$
GLUCOSE-BERT-Large	$\mathbf{54.7}_{(0.02)}$
RoBERTa-Large (baseline)	$56.5_{(0.06)}$
ATOMIC-RoBERTa-Large	
- Event only	$64.2_{(0.04)}$
- Event, Physical, Social	$61.8_{(0.04)}$
GLUCOSE-RoBERTa-Large	$\mathbf{66.1}_{(0.03)}$

Table 5.9: BCOPA-CE Accuracy results. Base model in *b-l-** is BERT-Large.

5.7.1 Error analysis

We dug a little deeper to understand how the augmented knowledge did or did not help achieve better results on our benchmarks. It is important to note that knowing what exact data points from ATOMIC$^{20}_{20}$ and GLUCOSE contributed to performance improvements is hard and may need a more rigorous analysis. However, it is still helpful to investigate the overlap between the data from these resources and our benchmarks' samples. To achieve that, in each benchmark, we picked our best-performing model and the baseline on the model with the same type (e.g., if our best-performing model is RoBERTa. Then we compared it to the vanilla RoBERTa baseline and not the BERT baseline.) Then for each model, we separated all samples in the test set that were correctly predicted across *all* the random seed runs to ensure the predictions were consistent and not random or by chance. Then we created a set of correctly predicted samples by our best model that our baseline failed to predict correctly. Moreover, for each sample in that set, we measured its semantic similarity with all data points in ATOMIC$^{20}_{20}$ and GLUCOSE. To measure the semantic overlap, we leveraged the `Sentence Transformers` (Reimers & Gurevych,

2019). [12] After computing the embeddings of samples,[13] we computed the cosine similarity between pairs of embeddings and separated pairs with at least 50% similarity. Our idea was that if we had a data point in ATOMIC$^{20}_{20}$ or GLUCOSE that has a high semantic similarly—concerning the interactions between events—with a data point in the benchmark, that semantic similarity might have contributed to the augmented model's performance improvement.

Table 5.10 shows examples of correctly classified samples with high context similarity with an entry in GLUCOSE. Out of 70,730 training samples in GLUCOSE, there are 3,588 and 253 pairs with 0.5 and 0.6 cosine similarity with a sample in COPA, respectively. As can be seen, there is a meaningful similarity in concept between these samples. For instance, from an entry in GLUCOSE, we know that *noticing an unfamiliar car* will result in a *curiosity feeling*. A similar question is asked in COPA about the result or effect of seeing *an unfamiliar car parked outside someone's house*. The correct answer to that question in COPA is *being suspicious* which is semantically similar to *curiosity feeling* which is a piece of knowledge we know from GLUCOSE as a result of *noticing an unfamiliar car*. This suggests that a model may have learned the relation between *seeing an unfamiliar object* and *a curiosity feeling* at the time of continual pretraining, which helped it later predict the correct answer when two similar events are involved. It is worth emphasizing that we may not be able to claim that this context similarity in training data is the real cause for the performance enhancement of our augmented models. However, it is still interesting to see that feeding a model with explicit causal statements helps the model express its causal knowledge.

[12] https://github.com/UKPLab/sentence-transformers

[13] The model we use is available on Huggingface: `sentence-transformers/all-mpnet-base-v2`

COPA Test Sample	GLUCOSE Similar Entry
The family went to the zoo. The children admired the animals. **(ask-for=result)**	The kids are excited to see they are at the zoo because the kids like(s) the zoo.
The phone rang. The man picked up the phone. **(ask-for=result)**	The guy answers the phone because the phone is ringing.
The trash bag was full. I took it to the dumpster. **(ask-for=result)**	I pick up the bag since the trash bag is full.
The runner sensed his competitor gaining on him. He sped up his pace. **(ask-for=result)**	Sam ran as fast as he could since sam feel(s) competitive.
The man got out of the shower. The hot water was gone. **(ask-for=cause)**	All the hot water is gone because my wife just used the shower.
The criminal was executed. He was convicted of murder. **(ask-for=cause)**	The judge convicts him because he is guilty.
The boy's forehead felt hot. His mother took his temperature. **(ask-for=result)**	Sean's mom takes his temperature caused sean's mom finds out he has a fever.
The fish bit the line. The fisherman reeled in the fish. **(ask-for=result)**	A huge fish gets on the line. As a result bob has a bite.
The man went to the doctor. The man felt ill. **(ask-for=cause)**	Tom goes to the doctor because tom feel(s) sick.
An unfamiliar car parked outside my house. I became suspicious. **(ask-for=result)**	I notice an unfamiliar car. As a result I feel(s) curiosity.

Table 5.10: Correctly classified samples in COPA paired with most semantically similar entries in GLUCOSE.

5.8 Conclusion

We introduced a simple framework for augmenting PLMs with commonsense knowledge that was created by automatically verbalizing ATOMIC$^{20}_{20}$ and GLUCOSE. Our results show that commonsense knowledge-augmented PLMs outperform traditional PLMs in cause–effect pair classification and answering commonsense causal reasoning questions. As the next step, it would be interesting to see how the previously proposed model improvement

methods or unbiased fine-tuning datasets can potentially enhance the performance of our knowledge-augmented models.

References

Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2021a). Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3554–3565).

Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2021b, June). Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3554–3565). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.278 doi: 10.18653/v1/2021.naacl-main.278

Arfé, B., & Boscolo, P. (2006). Causal coherence in deaf and hearing students' written narratives. *Discourse Processes*, *42*(3), 271–300.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on computational linguistics-volume 1* (pp. 86–90).

Bethard, S., Corvey, W. J., Klingenstein, S., & Martin, J. H. (2008). Building a corpus of temporal-causal structure. In *Lrec*.

Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 267–275.

Blanco, E., Castell, N., & Moldovan, D. I. (2008). Causal relation extraction. In *Lrec*.

Broniatowski, D. A., Hilyard, K. M., & Dredze, M. (2016). Effective vaccine communication during the disneyland measles outbreak. *Vaccine*, *34*(28), 3225–3228.

Broniatowski, D. A., & Reyna, V. F. (2017, February). *Combating Misinformation and Disinformation Online: The "Battle of the Narrative"* (White Paper). Washington,

D.C: Submitted to the Committee on a Decadal Survey of Social and Behavioral Sciences and Applications to National Security, National Academies of Sciences, Engineering, and Medicine.

Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 161–164).

Caselli, T., & Vossen, P. (2016). The storyline annotation and representation scheme (star): A proposal. In *Proceedings of the 2nd workshop on computing news storylines (cns 2016)* (pp. 67–72).

Caselli, T., & Vossen, P. (2017). The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the events and stories in the news workshop* (pp. 77–86).

Chang, D.-S., & Choi, K.-S. (2004). Causal relation extraction using cue phrase and lexical pair probabilities. In *International conference on natural language processing* (pp. 61–70).

Chang, D.-S., & Choi, K.-S. (2006). Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information processing & management*, *42*(3), 662–678.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, *48*(4), 1227–1237.

Cybulska, A., & Vossen, P. (2014). Guidelines for ecb+ annotation of events and their coreference. In *Technical report.* Technical Report NWR-2014-1, VU University Amsterdam.

Dalal, D., Arcan, M., & Buitelaar, P. (2021). Enhancing multiple-choice question answering

with causal knowledge. In *Proceedings of deep learning inside out (deelio): The 2nd workshop on knowledge extraction and integration for deep learning architectures* (pp. 70–80).

Dandignac, M., & Wolfe, C. R. (2020). Gist inference scores predict gist memory for authentic patient education cancer texts. *Patient Education and Counseling*, *103*(8), 1562–1567.

Dasgupta, T., Saha, R., Dey, L., & Naskar, A. (2018). Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th annual sigdial meeting on discourse and dialogue* (pp. 306–316).

Davison, J., Feldman, J., & Rush, A. M. (2019). Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1173–1178).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diehl, J. J., Bennetto, L., & Young, E. C. (2006). Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of abnormal child psychology*, *34*(1), 83–98.

Do, Q. X., Chan, Y. S., & Roth, D. (2011). Minimally supervised event causality identification. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 294–303).

Dredze, M., Broniatowski, D. A., Smith, M. C., & Hilyard, K. M. (2016). Understanding vaccine refusal: why we need social media now. *American journal of preventive medicine*, *50*(4), 550–552.

Drury, B., Oliveira, H. G., & de Andrade Lopes, A. (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering*, *28*(3), 361–400.

Duari, S., & Bhatnagar, V. (2021). Ffcd: A fast-and-frugal coherence detection method.

IEEE Access.

Dunietz, J., Carbonell, J. G., & Levin, L. (2018). Deepcx: A transition-based approach for shallow semantic parsing with complex constructional triggers. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1691–1701).

Dunietz, J., Levin, L., & Carbonell, J. (2015). Annotating causal language using corpus lexicography of constructions. In *Proceedings of the 9th linguistic annotation workshop* (pp. 188–196).

Dunietz, J., Levin, L., & Carbonell, J. (2017a). The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th linguistic annotation workshop* (pp. 95–104).

Dunietz, J., Levin, L., & Carbonell, J. (2017b, April). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th linguistic annotation workshop* (pp. 95–104). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-0812 doi: 10.18653/v1/W17-0812

Facebook. (2019). *Fact-checking on facebook: What publishers should know.* http://bit.ly/2ouEDwJ. ([Online])

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307.

Funke, D. (2019). *Facebook's fact-checking project sometimes forces hoaxers to change their content.* http://bit.ly/2m89KNr. ([Online])

Gao, L., Choubey, P. K., & Huang, R. (2019). Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 1808–1817).

Garcia, D., et al. (1997). Coatis, an nlp system to locate expressions of actions connected by

causality links. In *International conference on knowledge engineering and knowledge management* (pp. 347–352).

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the acl 2003 workshop on multilingual summarization and question answering-volume 12* (pp. 76–83).

Girju, R., Moldovan, D. I., et al. (2002). Text mining for causal relations. In *Flairs conference* (pp. 360–364).

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2007a). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th international workshop on semantic evaluations* (pp. 13–18). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1621474.1621477

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2007b, June). SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 13–18). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S07-1003

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review*, *23*(5), 1566–1575.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., ... others (2018). Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th acm conference on web science* (pp. 17–21).

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking reading comprehension*, 82–98.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments,*

& computers, *36*(2), 193–202.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Grysman, A., & Hudson, J. A. (2010). Abstracting and extracting: Causal coherence and the development of the life story. *Memory*, *18*(6), 565–580.

Guan, J., Huang, F., Zhao, Z., Zhu, X., & Huang, M. (2020). A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, *8*, 93–108.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360).

Han, M., & Wang, Y. (2021, August). Doing good or doing right? exploring the weakness of commonsense causal reasoning models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 151–157). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-short.20 doi: 10.18653/v1/2021.acl-short.20

Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J.-H., & Kazama, J. (2012). Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 619–630).

Hashimoto, C., Torisawa, K., Kloetzer, J., & Oh, J.-H. (2015). Generating event causality hypotheses through semantic relations. In *Twenty-ninth aaai conference on artificial intelligence.*

Hashimoto, C., Torisawa, K., Kloetzer, J., Sano, M., Varga, I., Oh, J.-H., & Kidawara, Y. (2014). Toward future scenario generation: Extracting event causality exploiting

semantic relation, context, and association features. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 987–997).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., ... Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the workshop on semantic evaluations: Recent achievements and future directions* (pp. 94–99). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1621969.1621986

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., ... Szpakowicz, S. (2010, July). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 33–38). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S10-1006

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., ... Szpakowicz, S. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 33–38).

Hidey, C., & McKeown, K. (2016a). Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 1424–1433).

Hidey, C., & McKeown, K. (2016b, August). Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1424–1433). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://

aclanthology.org/P16-1135 doi: 10.18653/v1/P16-1135

Hosseini, P., Broniatowski, D. A., & Diab, M. (2022, May). Knowledge-augmented language models for cause-effect relation classification. In *Proceedings of the first workshop on commonsense representation and reasoning (csrr 2022)* (pp. 43–48). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.csrr-1.6

Hosseini, P., Diab, M., & Broniatowski, D. A. (2019). Does causal coherence predict online spread of social media? In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (pp. 184–193).

Hosseini, P., Wolfe, C., Diab, M. T., & Broniatowski, D. A. (2022). Gispy: A tool for measuring gist inference score in text. In *Naacl 2022 workshop on narrative understanding*.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2020). Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2021). Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Aaai*.

Joskowicz, L., Ksiezyck, T., & Grishman, R. (1989). Deep domain models for discourse analysis. In *[1989] proceedings. the annual ai systems in government conference* (pp. 195–200).

Kang, D., Gangal, V., Lu, A., Chen, Z., & Hovy, E. (2017). Detecting and explaining causes from text for a time series event. *arXiv preprint arXiv:1707.08852*.

Kaplan, R. M., & Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal

relationships in text. *Knowledge Acquisition*, *3*(3), 317–337.

Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., & Inui, K. (2019). When choosing plausible alternatives, clever hans can be clever. *EMNLP 2019*, 33.

Khoo, C. S. (1995). *Automatic identification of causal relations in text and their use for improving precision in information retrieval* (Unpublished doctoral dissertation).

Khoo, C. S., Chan, S., & Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting on association for computational linguistics* (pp. 336–343).

Khoo, C. S., Kornfilt, J., Oddy, R. N., & Myaeng, S. H. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, *13*(4), 177–186.

King, T. (2020). Percent of your data will be unstructured in five years. *Retrieved February*, *16*, 80.

Knaebel, R. (2021, November). discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd workshop on computational approaches to discourse* (pp. 128–133). Punta Cana, Dominican Republic and Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.codi-main.12 doi: 10.18653/v1/2021.codi-main.12

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An erp study. *Journal of cognitive neuroscience*, *23*(5), 1230–1246.

Laban, P., Dai, L., Bandarkar, L., & Hearst, M. A. (2021). Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 1058–1064).

Lapata, M., Barzilay, R., et al. (2005). Automatic evaluation of text coherence: Models and representations. In *Ijcai* (Vol. 5, pp. 1085–1090).

Levi, O., Hosseini, P., Diab, M., & Broniatowski, D. (2019, November). Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In *Proceedings of the second workshop on natural language processing for internet freedom: Censorship, disinformation, and propaganda* (pp. 31–35). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-5004

Li, P., & Mao, K. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, *115*, 512–523.

Li, Z., Chen, T., & Van Durme, B. (2019). Learning to rank for plausible plausibility. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4818–4823).

Li, Z., Ding, X., Liu, T., Hu, J. E., & Van Durme, B. (2020). Guided generation of cause and effect..

Lin, Z., Ng, H. T., & Kan, M.-Y. (2011, June). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 997–1006). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P11-1100

Liu, S., Zeng, S., & Li, S. (2020). Evaluating text coherence at sentence and paragraph levels. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1695–1703).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ljubešić, N., Fišer, D., & Peti-Stantić, A. (2018, July). Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the third workshop on representation learning for NLP* (pp. 217–222).

Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-3028 doi: 10.18653/v1/W18-3028

Low, B.-T., Chan, K., Choi, L.-L., Chin, M.-Y., & Lay, S.-L. (2001). Semantic expectation-based causation knowledge extraction: A study on hong kong stock movement analysis. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 114–123).

Luo, Z., Sha, Y., Zhu, K. Q., Hwang, S.-w., & Wang, Z. (2016). Commonsense causal reasoning between short texts. In *Fifteenth international conference on the principles of knowledge representation and reasoning.*

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272).

Mirza, P. (2014). Extracting temporal and causal relations between events. In *Proceedings of the acl 2014 student research workshop* (pp. 10–17).

Mirza, P., Sprugnoli, R., Tonelli, S., & Speranza, M. (2014). Annotating causality in the tempeval-3 corpus. In *Eacl 2014 workshop on computational approaches to causality in language (catocl)* (pp. 10–19).

Mirza, P., & Tonelli, S. (2014). An analysis of causality between events and its relation to temporal information. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 2097–2106).

Mirza, P., & Tonelli, S. (2016). Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 64–75).

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., & Vanderwende, L. (2016). Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the fourth workshop on events* (pp. 51–61).

Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O., & Chu-Carroll, J. (2020). Glucose: Generalized and contextualized story explanations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4569–4586).

Mulder, G., & Sanders, T. J. (2012). Causal coherence relations and levels of discourse representation. *Discourse Processes*, *49*(6), 501–522.

Ning, Q., Feng, Z., Wu, H., & Roth, D. (2018a). Joint reasoning for temporal and causal relations. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2278–2288).

Ning, Q., Feng, Z., Wu, H., & Roth, D. (2018b, July). Joint reasoning for temporal and causal relations. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2278–2288). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-1212

O'Gorman, T., Wright-Bettner, K., & Palmer, M. (2016). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd workshop on computing news storylines (cns 2016)* (pp. 47–56).

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*.

Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Phu, M. T., & Nguyen, T. H. (2021). Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3480–3490).

Pitler, E., & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 13–16).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The penn discourse treebank 2.0. In *Lrec*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Reese, E., Haden, C. A., Baker-Ward, L., Bauer, P., Fivush, R., & Ornstein, P. A. (2011). Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of Cognition and Development*, *12*(4), 424–462.

Reganti, A. N., Maheshwari, T., Kumar, U., Das, A., & Bajpai, R. (2016). Modeling satire in english text for automatic detection. In *2016 ieee 16th international conference on data mining workshops (icdmw)* (pp. 970–977).

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical decision making*, *28*(6), 850–865.

Reyna, V. F. (2012a). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*.

Reyna, V. F. (2012b). Risk perception and communication in vaccination decisions: A fuzzy-trace theory approach. *Vaccine*, *30*(25), 3790–3797.

Reyna, V. F. (2021). A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *Proceedings of the National Academy of Sciences*, *118*(15), e1912441117.

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 aaai spring symposium series.*

Rogers, A., Gardner, M., & Augenstein, I. (2021). Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17).

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019, November). Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international*

joint conference on natural language processing (emnlp-ijcnlp) (pp. 4463–4473). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1454 doi: 10.18653/v1/D19-1454

Selfridge, M., Daniell, J., & Simmons, D. (1985). Learning causal models by understanding real-world natural language explanations. *The engineering of knowledge-based systems*, 378–383.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, *9*(1), 4787.

Sharp, R., Pyarelal, A., Gyori, B. M., Alcock, K., Laparra, E., Valenzuela-Escárcega, M. A., ... others (2019). Eidos, indra, & delphi: From free text to executable causal models. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Demonstrations. association for computational linguistics.*

Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 43–65). Springer.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Shu, K., Wang, S., & Liu, H. (2017). Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

Sorgente, A., Vettigli, G., & Mele, F. (2013). Automatic extraction of cause-effect relations in natural language text. *DART@ AI* IA*, *2013*, 37–48.

Tapanainen, P., & Jarvinen, T. (1997). A non-projective dependency parser. In *Fifth conference on applied natural language processing* (pp. 64–71).

Trabasso, T., et al. (1982). Causal cohesion and story coherence.

Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and language*, *24*(5), 595–611.

Tsur, O., & Rappoport, A. (2012). What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth acm international conference on web search and data mining* (pp. 643–652).

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second joint conference on lexical and computational semantics (* sem), volume 2: Proceedings of the seventh international workshop on semantic evaluation (semeval 2013)* (Vol. 2, pp. 1–9).

Van den Broek, P. (1990). Causal inferences and the comprehension of narrative texts. In *Psychology of learning and motivation* (Vol. 25, pp. 175–196). Elsevier.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6–10.

Wolfe, C. R., Dandignac, M., & Reyna, V. F. (2019). A theoretically motivated method for automatically evaluating texts for gist inferences. *Behavior research methods*, *51*(6), 2419–2437.

Wolfe, C. R., Dandignac, M., Wang, C., & Lowe, S. R. (2021). Gist inference scores predict cloze comprehension "in your own words" for native, not esl readers. *Health Communication*, 1–8.

Wolfe, C. R., Widmer, C. L., Torrese, C. V., & Dandignac, M. (2018). A method for automatically analyzing intelligent tutoring system dialogues with coh-metrix. *Journal of Learning Analytics*, *5*(3), 222–234.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General*, *136*(1), 82.

Wu, S., & He, Y. (2019). Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 2361–2364).

Yang, J., Han, S. C., & Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 1–26.

Yu, B., Li, Y., & Wang, J. (2019). Detecting causal language use in science findings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4656–4666).

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 93–104).