EXPLANATION-BASED SCIENTIFIC NATURAL LANGUAGE INFERENCE

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2022

By
Marco Valentino
Department of Computer Science

Contents

List of Tables

List of Figures

Abstract

EXPLANATION-BASED SCIENTIFIC NATURAL LANGUAGE INFERENCE
Marco Valentino
A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2022

Building systems that can explain and understand the world is a long-standing goal for Artificial Intelligence (AI). The ability to explain, in fact, constitutes an archetypal feature of human rationality, underpinning communication, learning, and generalisation, as well as one of the mediums enabling scientific discovery through the formulation of explanatory theories. As part of this long-term goal for AI, a large body of research in Natural Language Processing (NLP) focuses on the development and evaluation of explanation-based inference models, capable of reasoning through the interpretation and generation of *natural language explanations*.

However, research in Explanation-based Natural Language Inference (NLI) presents several fundamental challenges. Firstly, the applied methodologies are still poorly informed by theories and accounts of explanations. Current work, in fact, rarely recur to formal characterisations of the nature and function of explanations, and are limited to generic explanatory properties. This gap between theory and practice poses the risk of slowing down progress in the field, missing the opportunity to formulate clearer hypotheses on inferential properties of explanations and well-defined evaluation methodologies. Secondly, Explanation-based NLI models still lack robustness and scalability for real-world applications. In particular, existing approaches suffer from several limitations when it comes to composing explanatory inference chains from large facts banks and performing abstraction for NLI in complex domains.

This thesis focuses on *scientific explanation* as a rich theoretical and experimental framework for advancing research in Explanation-based NLI. In particular, the goal of the thesis is to investigate some of the fundamental challenges in the field from both a theoretical and an empirical perspective, attempting to derive a grounded epistemological-linguistic characterisation to inform the construction of more accurate and scalable Explanation-based NLI models in the scientific domain.

Overall, the research described in the thesis can be summarised in the following scientific contributions:

1. An extensive study on the notion of a scientific explanation from both a categorical and a corpus-based perspective aimed at deriving a grounded characterisation for explanation-based NLI. The study reveals that explanations cannot be entirely defined in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*, fitting the event to be explained into a broader underlying regularity. Moreover, the study suggests that unification is an intrinsic property of existing corpora, emerging as explicit and recurring explanatory patterns in natural language.

2. A novel computational model based on the notion of *explanatory power* as defined in the unificationist account of scientific explanation. Specifically, the model can be adopted to capture explicit *explanatory patterns* emerging in corpora of natural language explanations and flexibly integrated into explanation-based NLI architectures for downstream inference tasks.

3. An empirical study on the impact of the explanatory power model on explanation-based NLI in the scientific domain, integrating it within sparse, dense and hybrid architectures, and performing a comprehensive evaluation in terms of *inferential properties*, *accuracy* and *scalability*.

4. A systematic evaluation methodology to inspect and verify the *logical properties* of explanation-supporting corpora and benchmarks. The study, aimed at providing a critical quality assessment of gold standards for NLI, reveals that a majority of human-annotated explanations represent invalid arguments, ranging from being incomplete to containing identifiable logical errors.

Declaration

i. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

ii. The material presented in this thesis represents the candidate's own work except where stated otherwise.

Marco Valentino

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.manchester.ac.uk/library/aboutus/regulations`) and in The University's policy on presentation of Theses

Acknowledgements

Undertaking a PhD has been one of the most intense, inspiring, and exciting journeys of my entire life. This achievement would not have been possible without the support of all the amazing people I have met along the way.

Special thanks to my supervisor Andre Freitas for providing me with this incredible opportunity. His constant encouragement to go beyond my limits and precious advice have been fundamental in shaping this work and my qualities as a researcher.

I would like to thank my examiners Louise Dennis and Jochen Leidner for thoroughly reviewing this thesis and for the profound and valuable discussion that emerged during the viva.

I was extremely lucky to share my PhD with real friends. Thanks to Mohan and Deborah for their constant presence in both good and bad times. Most of the memories I have about my PhD involve you both, and I will never forget your support throughout this journey. Many thanks also to Jake, Alber, and Julia for enriching my PhD with fun, joy, and good memories. All of you have made me feel like I have found a new family, and I am sure our friendship will remain forever.

I would like to thank all the people from the Reasoning & Explainable AI lab and the Department of Computer Science I had the opportunity to meet and work with, including but not limited to Mauricio, Conor, Hanadi, Mingyang, Mael, Crystal, Viktor, Gian, Mario, Martina, Ruba, Guy, Oskar, Zili, Jordan, Edoardo, Danilo.

Special thanks to my family for always believing in my choices, and providing all the support needed to embark on this journey. Leaving my country was one of the hardest decisions ever made, but thanks to your constant presence I have never felt far from home. I hope you are proud of this achievement as much as I am.

Finally, I am incredibly grateful to Francesca, the love of my entire life. You are the real witness to the hard work and passion I put into achieving this milestone, and I couldn't have wished for a different person to be there by my side. You may not fully realise it, but if I am a better person today, it is mostly because of your support.

Chapter 1

Introduction

1.1 Motivation

Building systems that can explain and understand the world is a long-standing goal for *Artificial Intelligence (AI)* [113, 116, 149]. The ability to explain, in fact, constitutes an archetypal feature of human rationality, underpinning communication, learning, and generalisation, as well as one of the mediums enabling scientific discovery and progress through the formulation of explanatory theories [109, 131, 91, 38].

While explanation appears to be a fundamental component of human intelligence, the dominant paradigm in AI-related fields such as *Natural Language Inference (NLI)* is currently represented by Deep Learning architectures [163, 39], whose general inference framework relies on end-to-end predictive power without supporting explanations. However, despite NLI systems based on Deep Learning demonstrated remarkable performance in specific benchmarks [13, 167], an increasing amount of empirical evidence suggests that end-to-end architectures do not actually learn the underlying rules and principles of the task at hand, but rather rely on superficial annotation artifacts and biases, being susceptible to shortcuts learning and unable to generalise to out-of-distribution examples [49, 101, 141, 137]. Moreover, end-to-end predictive models are generally regarded as black-boxes, whose lack of interpretability to the end-user poses serious concerns in terms of applicability and trust for real-world scenarios [51, 12].

These limitations have led to reconsider the role that explanation plays for learning and inference with natural language [15, 187, 123, 72]. In contrast to the existing end-to-end paradigm, an emerging line of research in NLI focuses on the development and evaluation of *explanation-based inference* models, whose goal is to perform predictions through the explicit construction of a *natural language explanation* [31, 67, 173, 143]. In this context, explanation might constitute a potential way to mitigate some of the well-known limitations in the field, providing a mechanism for learning explicit, interpretable, and generalisable reasoning strategies [151].

Research in Explanation-based NLI, however, presents several fundamental challenges. Firstly, the applied methodologies are still poorly informed by theories and accounts of explanations [131, 178]. This gap between theory and practice poses the risk of slowing down progress, missing the opportunity to formulate clearer hypotheses on

the inferential properties of natural language explanations and well-defined evaluation methodologies [16, 69]. Secondly, Explanation-based NLI models still lack robustness and scalability for real-world applications. In particular, existing approaches suffer from several limitations when it comes to composing explanatory inference chains from large fact banks and performing abstraction for NLI in complex domains [82, 46, 71].

1.1.1 Scientific Explanation

This thesis focuses on *scientific explanation* as a rich framework for advancing research in Explanation-based NLI. The motivation behind the focus on the scientific domain is two-fold:

1. Although scientific explanation remains a complex epistemological subject, the quest for delivering a rigorous account of its nature and function has produced a set of quasi-formal models that clarify to some extent the nature of the concept [131, 53]. These epistemological accounts can provide fundamental insights for the construction of Explanation-based NLI systems, contributing to reduce the gap between theory and practice.

2. Scientific explanations posses specific features that are particularly challenging for Explanation-based NLI models. In particular, a scientific explanation typically derives the occurrence of observable phenomena from underlying mechanisms and hidden regularities, a feature that often requires abstraction through the integration of multiple inference steps [67, 72]. These features can provide an experimental framework for testing the limitations of existing systems, informing the development of novel inference strategies and evaluation methodologies.

Therefore, the scientific domain allows investigating some of the fundamental challenges in the field of Explanation-based NLI from both a theoretical and an empirical perspective. Specifically, the main objective of the thesis it to attempt to derive a grounded epistemological-linguistic characterisation of natural language explanations for the construction of more accurate and scalable explanation-based inference models.

1.2 Background

1.2.1 Problem Definition

Given a certain reasoning problem expressed in natural language, *Explanation-based NLI* aims at automatically finding the correct solution through the construction of a *natural language explanation*. In that regard, the construction of an explanation represents the central mechanism and process through which an Explanation-based NLI system performs predictions.

Thanks to its general definition, Explanation-based NLI can be applied to model a vast range of problems in different domains, spanning from Question Answering (QA)

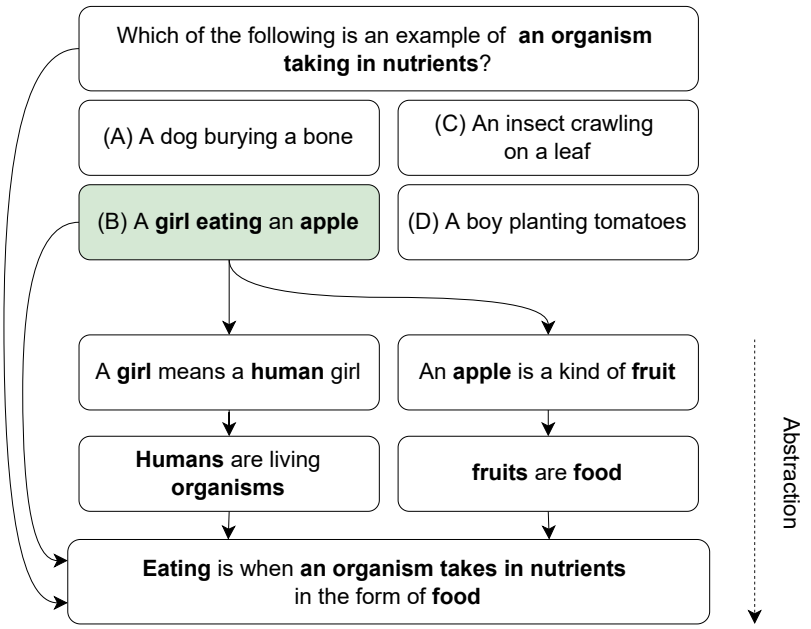


Figure 1.1: An example of Explanation-based NLI for answering a multiple-choice science question from [66].

[179, 187], to Textual Entailment (TE) [15] and Fact Verification (FV) [5]. Figure 1.1 illustrates a specific example of an explanation-based inference for answering a multiple-choice science question [66, 72].

In general, an explanation for a NLI problem is defined as a set of interconnected sentences supporting the final answer. In this context, the construction of an explanation typically requires multiple inference steps to retrieve and compose sentences from external knowledge sources [67]. As shown in the example illustrated in figure 1.1, multiple inference steps are often necessary to address unseen problems in complex domains since it is unlikely to find a single, contiguous passage of text containing a valid and complete explanation supporting the correct answer. The problem of retrieving and aggregating multiple supporting facts for Explanation-based NLI is generally studied under the name of *multi-hop inference* [187, 87, 180].

The type and structure of multi-hop inference can vary according to the specific reasoning task and the domain under consideration. Scientific explanations, in particular, require the articulation and integration of commonsense and scientific knowledge along with the encoding of abstraction and grounding mechanisms (e.g., *"humans are living organisms"*, *"fruits are food"*) for the identification of relevant explanatory statements (e.g., *"eating is when an organism takes in food"*), a feature that makes multi-hop inference particularly challenging [22, 86, 161, 152].

This work focuses on scientific NLI problems, investigating two tasks at the core of Explanation-based NLI, namely *Explanation Regeneration* [71] and *Abductive Natural*

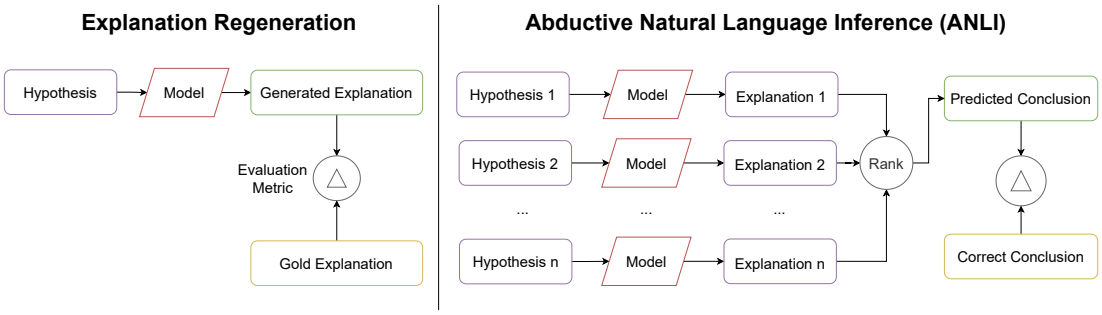


Figure 1.2: High-level, schematic representation of the tasks of Explanation Regeneration and Abductive Natural Language Inference (ANLI). Explanation Regeneration consists in reconstructing the gold explanation supporting a *true* hypothesis provided as input, while ANLI consists in predicting the most plausible conclusion among a set of mutually exclusive hypotheses (i.e., only one known to be true) through an *inference to the best explanation*.

Language Inference (ANLI) [10]. Here, we provide a formalisation of the specific tasks explored throughout the thesis. A high-level, schematic representation of the tasks is illustrated in Figure 1.2.

Explanation Regeneration

Given a statement expressed in natural language, known as hypothesis h, the task of *Explanation Regeneration* consists in reconstructing the ground-truth (gold) explanation supporting h by retrieving and composing a sequence of atomic sentences $E_{seq} = f_1, \ldots, f_n$ from an external fact bank (Fig 1.2, left).

The main focus of Explanation Regeneration is to evaluate the quality of the explanations generated by a given model, typically quantifying the similarity/difference between automatically generated explanations and human-annotated explanations. For this reason, Explanation Regeneration is regarded as a crucial intermediate task for the development and evaluation of Explanation-based NLI systems.

However, it is important to notice that the aim of the task is not to evaluate end-to-end inference performance since the input hypothesis is typically represented by a true statement containing the correct answer. For example, in the context of multiple-choice question answering, the input hypothesis is derived from converting the question and the correct choice into an affirmative statement (e.g., *"A girl eating an apple is an example of an organism taking in nutrients"* in Figure 1.1).

Deduction	Induction	Abduction
All the beans in this bag are white These beans are from this bag	These beans are from this bag These beans are white	All the beans in this bag are white These beans are white
These beans are white	All the beans in this bag are white	These beans are from this bag

Table 1.1: The three types of reasoning identified by C. S. Peirce [121].

Abductive Natural Language Inference (ANLI)

The notion of abduction or abductive inference has been the subject of a long discussion in several fields including Philosophy, Logic and Computer Science [23, 54, 61, 1, 106, 166]. The term was introduced by C. S. Peirce [121] to describe a form of human reasoning that differs from both induction (i.e., generalisation from specific examples to rules) and deduction (derivation of conclusions from general rules). Specifically, abductive inference is typically defined as the process of finding the most likely explanation for an observation or a set of observations (see example in Table 1.1). Differently from deduction, abductive inference cannot guarantee the correctness of the final conclusion. This is because, in general, more than one plausible explanation exists for a given set of observations, and the correct explanation cannot be deductively derived from the rules. Therefore, abductive inference typically relies on a set of heuristics to derive the most plausible conclusion.

While the term has been historically used with slighly different meanings, this thesis refers to abduction as *an inference to the best explanation* [54, 106], intended as the process of selecting the most plausible explanation among competing ones. Specifically, for a given set of alternative, mutually exclusive natural language hypotheses $H = \{h_1, h_2, \ldots, h_n\}$, we define *Abductive Natural Language Inference (ANLI)* as the task of selecting the hypothesis in H that is supported by the *best explanation* (Fig 1.2, right).

Differently from Explanation Regeneration, the focus of ANLI is to adopt explanation as a mechanism to perform end-to-end inference and predict the solution to a NLI problem. The evaluation, therefore, generally concentrates on comparing the final conclusion of the system with the expected solution. ANLI can be regarded as complementary to Explanation Regeneration for building and evaluating Explanation-based NLI systems as the same inference model can be adopted to construct an explanation for each hypothesis in H and subsequently predict the final answer. Therefore, ANLI systems typically subsumes Explanation Regeneration models, extending them with an

additional component for selecting the best explanation among competing ones.

The high-level schema of ANLI can be instantiated in different ways according to the specific task at hand. For instance, in case the task to solve is a multiple-choice question answering problem, each hypothesis will represent a possible candidate answer. On the other hand, for different tasks such as Fact Verification, H might include a single hypothesis, with the final goal of predicting whether the given hypothesis is supported or rejected by the generated explanation.

1.2.2 Problem Statement

Here, we present a set of challenges and limitations associated with state-of-the-art models and evaluation methodologies for Explanation-based NLI which will represent the main focus of the thesis.

Epistemological-Linguistic Perspective

Current lines of research in Explanation-based NLI mainly focus on the development of explanation-based inference models [41, 32, 151] and explanation-centred resources for the evaluation [173, 31, 179, 74, 72, 154].

However, the applied methodologies are still poorly informed by epistemological and linguistic accounts on the nature and function of explanations [178, 113, 147]. When describing natural language explanations for model and benchmark creation, existing work rarely recurs to formal characterisations of what constitute a *valid explanatory argument*, and are often limited to the indication of generic properties in terms of *supporting evidence* or *entailment* [187, 15, 158, 31]. Bridging the gap between theory and practice, therefore, is necessary for enabling progress in the field and providing new opportunities to formulate clearer research hypotheses on *inferential properties* of explanations and well-defined evaluation methodologies [16, 158, 69, 24].

Multi-hop Inference and Explanatory Relevance

The problem of constructing a natural language explanation for a given hypothesis through multi-hop inference involves the capability of dynamically and correctly estimating the *explanatory relevance* of sentences in a fact bank. However, in complex domains such as scientific NLI, the problem of estimating explanatory relevance can be particularly challenging for the following reasons:

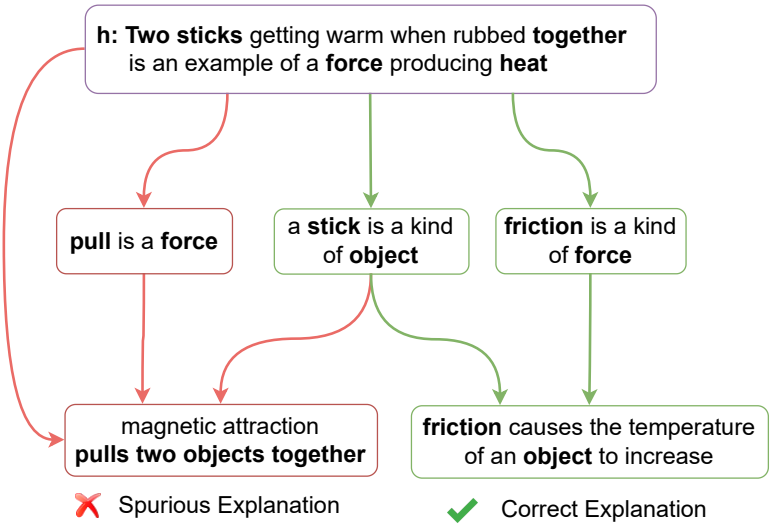


Figure 1.3: An example of semantic drift in multi-hop inference causing the construction of a spurious explanation.

1. **Hidden Explanation Structure:** The high-level structure of the explanation is generally not evident from the decomposition of the hypothesis, that is, the type of facts required for the inference cannot be derived from the surface form of the NLI problem to be solved;

2. **Abstract Explanatory Facts:** Core explanatory statements tend to be abstract and share a low number of terms with the hypothesis. This is particularly evident in scientific explanations, where observable phenomena (e.g., *"Two sticks getting warm when rubbed"*) are typically explained in terms of high-level regularities and underlying mechanisms (e.g., *"Friction causes the temperature of an object to increase"*);

3. **Distracting Information:** External fact banks usually contain a large amount of irrelevant sentences, a feature that lowers the probability of identifying relevant facts in large corpora and that can contribute to the generation of spurious explanations leading to wrong conclusions.

Some of these challenges are illustrated in the example in Figure 1.3.

While existing approaches attempt to address these challenges employing iterative and path-based methods [94], or explicit constraints to guide the generation of a plausible explanation graph [84], empirical studies have shown that multi-hop inference models are still not robust in complex domains and tend to suffer from a phenomenon known as *semantic drift* – i.e., the tendency of drifting out-of-context as the number of

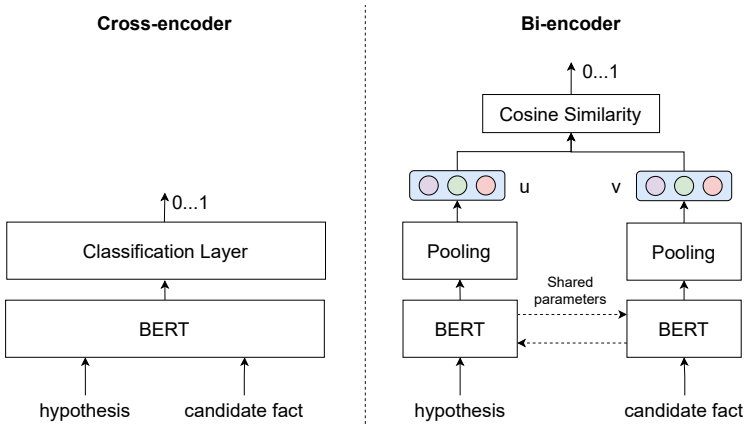


Figure 1.4: Different Transformer-based frameworks for estimating explanatory relevance. Cross-encoders (left) tend to be more robust thanks to the application of a classification mechanism. However, in contrast to bi-encoders (right), the representation vectors of candidate facts cannot be pre-computed and stored in apposite dense indexes for efficient inference [125].

inference steps increases [82, 46]. In other words, the challenges involved in estimating explanatory relevance in natural language induce a phenomenon of error propagation, which accumulates proportionally to the number of hops required to construct the final explanation.

Accuracy vs Scalability Trade-off

As a consequence of the challenges that multi-hop inference entails, the design of Explanation-based NLI systems in complex domains usually involves making a choice between accuracy and scalability.

State-of-the-art models typically focus on accuracy, leveraging the power of the self-attention mechanism in Transformers [39, 163]. These architectures, known as *cross-encoders* (Fig.1.4, left), are trained for sequence classification to estimate the relevance of candidate facts and compose valid inference chains [19, 34, 21, 6]. However, cross-encoders make multi-hop inference intrinsically inefficient and not scalable. These architectures, in fact, do not allow for the construction of dense indexes to cache the representation of candidate sentences in large fact banks, resulting in prohibitively slow inference time for real-world applications [63].

While more scalable and computationally efficient solutions exist (i.e, bi-encoders), their applicability for multi-hop inference in complex domain is still under-explored. The robustness and performance of scalable architectures, in fact, tend to be significantly

lower than state-of-the-art models. Therefore, additional research is required to bridge the gap and find a better trade-off.

1.3 Research Questions and Objectives

This thesis attempts to tackle part of the open research problems previously discussed. Specifically, the aim of this work is to investigate the nature and function of explanations from an epistemological-linguistic point of view to inform the construction of more accurate and scalable Explanation-based NLI systems in the scientific domain.

Given the general objective of the thesis, the main high-level research question can be formulated as follows:

- ***RQ0:*** *"Can specific epistemological-linguistic aspects of scientific explanations inform the construction of more accurate and scalable Explanation-based NLI models?"*

We further break down the overall objective of the thesis into more specific research questions that will be explored in details in each chapter. The following sections discuss the specific research questions along with a general outline of the methodology adopted to answer them.

1.3.1 Scientific Explanation and Natural Language

The first part of the thesis aims at exploring the notion of a scientific explanation from an epistemological-linguistic perspective, focusing on the nature and function of explanatory arguments in science:

- ***RQ1:*** *"What is the nature and function of an explanatory argument from an epistemological-linguistic perspective?"*

This question represents the first fundamental step to elaborate more specific hypotheses about the construction of Explanation-based NLI models. Specifically, we aim at addressing **RQ1** by systematically reviewing the main modern accounts of scientific explanation developed in Philosophy of Science. We hypothesise that understanding the main features of a scientific explanation from a theoretical perspective can provide fundamental insights on linguistic aspects emerging in natural language explanations. In particular, we are interested in understanding how the main features presented in

the accepted epistemological accounts manifest as explicit linguistic patterns in natural language explanations:

- ***RQ2:*** *"How do linguistic patterns emerge in natural language explanations?"*

This question will be addressed through a corpus analysis on explanation-centred resources and informed by the systematic survey of the epistemological accounts. Specifically, the corpus analysis will investigate emerging features of scientific explanations through a mixture of quantitative and qualitative methodologies. The main objective of **RQ2** is to derive concrete linguistic hypotheses for the subsequent construction of Explanation Regeneration and Abductive NLI models in the scientific domain.

1.3.2 Explanatory Patterns for Explanation-based NLI

After clarifying the nature of a scientific explanation from an epistemological-linguistic point of view, the thesis will focus on Explanation-based NLI. Specifically, we hypothesise that some form of explicit explanatory patterns revealed through the aforementioned study can be leveraged to support the development of Explanation-based NLI models:

- ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"*

In particular, we aim to answer **RQ3** by defining a computational model of explanatory relevance informed by linguistic patterns in explanation-centred resources, leveraging it for complex Explanation-based NLI problems. The computational model will then be evaluated on specific Explanation Regeneration and Abductive NLI tasks with a focus on measuring its impact on downstream accuracy, explanation quality, and semantic drift.

1.3.3 Hybrid Models for Accurate and Scalable Inference

A central objective of the thesis is to find a better trade-off between accuracy and scalability for Explanation-based NLI, proposing a new framework that could jointly optimise inference performance and computational efficiency. In that regard, we hypothesise that this goal can be achieved through hybrid architectures that integrate latent and explicit representational models:

- ***RQ4:*** *"Can hybrid models integrating latent and explicit representations provide a framework for a better accuracy-scalability trade-off in Explanation-based NLI?"*

To answer **RQ4**, we focus on bi-encoder architectures, which possess the property of being intrinsically scalable, investigating the impact of integrating Transformer-based representations with explicit models of explanatory relevance.

1.3.4 Inferential Properties in Explanations Gold Standards

Existing explanation-centred resources provide linguistic evidence on how humans construct explanations that are perceived as plausible, coherent and complete. However, while these resources are adopted to develop and evaluate explanation-based inference models, it is not yet clear what inferential properties they actually encode. An objective of the thesis is to better characterise the nature of these gold standards, providing insights for developing better evaluation methodologies in the field.

To this end, we hypothesise that human-annotated explanations represent valid and complete arguments from which the answer for a given NLI problem logically follows. Specifically, we aim at investigating the following research question:

- ***RQ5:*** *"Do natural language explanations in existing gold standards represent valid and complete logical arguments?"*

To answer **RQ5**, we aim at developing a systematic evaluation methodology to inspect and verify the *logical properties* of explanation-centred corpora and provide a critical assessment of existing gold standards for Explanation-based NLI.

1.4 Contribution

The research described in this thesis can be summarised in the following scientific contributions:

1. **C1:** An extensive study on the notion of a scientific explanation from both a categorical and a corpus-based perspective aimed at deriving a grounded characterisation for Explanation-based NLI. The study reveals that explanations cannot be entirely defined in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*, fitting the event to be explained into a broader

underlying regularity. Moreover, the study suggests that unification is an intrinsic property of existing corpora, emerging as explicit and recurring linguistic patterns in natural language explanations.

2. **C2:** A novel computational model of explanatory relevance based on the notion of *explanatory power* as defined in the unificationist account of scientific explanation. Specifically, we empirically demonstrate that the model can be adopted to capture explicit *explanatory patterns* emerging in corpora of natural language explanations and flexibly integrated in Explanation-based NLI architectures for downstream inference tasks.

3. **C3:** An empirical study on the impact of the explanatory power model on Explanation-based NLI in the scientific domain, integrating it within sparse, dense and hybrid architectures, and performing a comprehensive evaluation in terms of *inferential properties*, *accuracy* and *scalability* for Explanation Regeneration and Abductive NLI.

4. **C4:** A systematic evaluation methodology to inspect and verify the *logical properties* of explanation-supporting corpora and benchmarks. The study, aimed at providing a critical quality assessment of gold standards for NLI, reveals that a majority of human-annotated explanations represent invalid arguments, ranging from being incomplete to containing identifiable logical errors.

1.5 Thesis Outline

The thesis is organised as follows:

Chapter 2 investigates the notion of a scientific explanation from an epistemological and linguistic perspective combining a systematic survey of the modern accounts of scientific explanation with a corpus analysis of explanation-based corpora in the scientific domain. The aim of this chapter is to provide the foundations for the formulation of the research hypotheses explored in the remaining of the thesis.

Chapter 3 introduces a novel model of explanatory power for Explanation-based NLI informed by the notion of unification in scientific explanation. In particular, we empirically evaluate the model on Explanation Regeneration tasks in the scientific domain investigating its impact on accuracy and semantic drift for sparse models. Moreover, the model is combined with downstream Transformers to model abductive NLI in a multiple-choice QA setting.

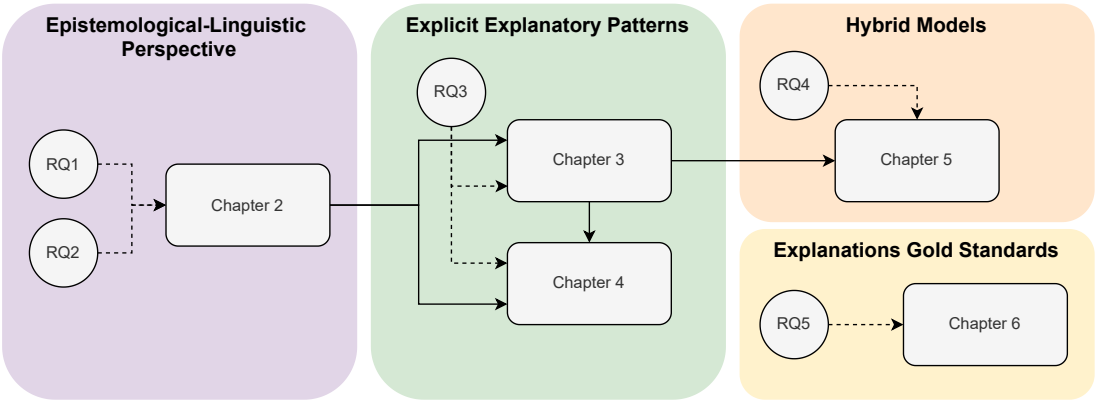


Figure 1.5: Overall structure of the thesis, research questions, and dependencies between the chapters.

Chapter 4 focuses on the impact of explicit explanatory patterns on abductive NLI for scientific and commonsense reasoning tasks. Specifically, the chapter integrates the notion of explanatory power in a case-based reasoning framework for abductive NLI, investigating the impact of retrieving and adapting explanations for similar cases solved in the past on unseen inference problems.

Chapter 5 further investigates the impact of the explanatory power model, extending the framework presented in Chapter 3 with dense models based on the bi-encoder architecture. Specifically, this chapter presents a hybrid autoregressive model for explanation-based inference, with a focus on the trade-off between accuracy and scalability in Explanation Regeneration.

Chapter 6 connects back to some of the research questions explored in Chapter 2, investigating the inferential properties of human-annotated explanations in explanation-centred resources for NLI. Specifically, the chapter analyses natural language explanations from an entailment perspective exploring their logical properties. To this end, we propose a systematic evaluation methodology which is subsequently applied to three popular Explanation-based NLI datasets.

Finally, **Chapter 7** revisits the main research questions and objectives, providing a discussion on the main findings, together with limitations and open research questions for future work.

1.6 Publications

The chapters presented in this thesis are based on the following publications:

- **Chapter 2:** Marco Valentino, André Freitas. *Scientific Explanation and Natural Language: A Unified Epistemological-Linguistic Perspective for Explaination-based NLI*. (Under Review) [157]. The thesis author designed the methodology presented in both survey and corpus analysis, leading the writing of the manuscript. André Freitas provided support for shaping and scoping the work, giving key comments and suggestions for the final revision.

- **Chapter 3:** Marco Valentino, Mokanarangan Thayaparan, André Freitas. *Unification-based Reconstruction of Multi-hop Explanations for Science Questions*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021) [161]. The thesis author designed the presented model of explanatory power, leading the writing of the manuscript. Mokanarangan Thayaparan provided fundamental support for experimental and methodological design, in particular for the fine-tuning of the transformer model adopted for question answering. André Freitas provided support and supervision throughout the publication process.

- **Chapter 4:** Marco Valentino, Mokanarangan Thayaparan, André Freitas. *Case-based Abductive Natural Language Inference*. (Under Review) [160]. The thesis author designed the case-based methodology, leading the writing of the manuscript. Mokanarangan Thayaparan provided conceptual and experimental support, helping shape the writing of the paper. André Freitas provided support and supervision throughout the publication process.

- **Chapter 5:** Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, André Freitas. *Hybrid Autoregressive Inference for Scalable Multi-Hop Explanation Regeneration*. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022, Oral Presentation) [159]. The thesis author designed the autoregressive inference model, leading the writing of the manuscript. Mokanarangan Thayaparan and Deborah Ferreira provided conceptual and experimental support, helping shape the writing of the paper. André Freitas provided support and supervision throughout the publication process.

- **Chapter 6:** Marco Valentino, Ian Pratt-Hartmann, André Freitas. *Do Natural Language Explanations Represent Valid Logical Arguments? Verifying Entailment in Explainable NLI Gold Standards*. Proceedings of the 14th International Conference on Computational Semantics (IWCS 2021) [158]. The thesis author

designed the verification methodology, leading the writing of the manuscript. Ian Pratt-Hartman provided fundamental support for the formalisation and annotation phase, also giving key suggestions for the writing. André Freitas provided support and supervision throughout the publication process.

Chapter 2

Scientific Explanation and Natural Language

This chapter investigates ***RQ1:*** *"What is the nature and function of an explanatory argument from an epistemological-linguistic perspective?"* and ***RQ2:*** *"How do linguistic patterns emerge in natural language explanations?"*. To this end, the chapter combines a systematic survey of the modern epistemological accounts of scientific explanation with a corpus analysis on natural language explanations, attempting to clarify the nature and function of explanatory arguments from both a top-down (categorical) and a bottom-up (corpus-based) perspective.

Through a mixture of quantitative and qualitative methodologies, the presented study allows deriving the following main conclusions: (1) Explanations cannot be entirely characterised in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*; (2) An explanation must cite *causes* and *mechanisms* that are responsible for the occurrence of the event to be explained; (3) While natural language explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms, also accounting for pragmatic elements such as *definitions*, *properties* and *taxonomic relations* (4) Patterns of *unification* naturally emerge in corpora of explanations even if not intentionally modelled; (5) Unification is realised through a process of *abstraction*, whose function is to provide the inference substrate for subsuming the event to be explained under recurring patterns and high-level regularities[1].

[1]This chapter follows the publication "Scientific Explanation and Natural Language: A Unified Epistemological-Linguistic Perspective for Explanation-based NLI" [157].

2.1 Introduction

Building models capable of performing complex inference through the generation of *natural language explanations* represents a fundamental research goal for Explanation-based NLI [41, 32, 151]. However, while current lines of research focus on the development of explanation-based models and benchmarks [173, 31, 179, 74, 72, 154], the applied methodologies are still poorly informed by formal accounts and discussions on the nature of explanation [178, 169, 113, 147]. When describing natural language explanations, in fact, existing work rarely recur to formal characterisations of what constitute a *valid explanatory argument*, and are often limited to the indication of generic properties in terms of *supporting evidence* or *entailment* [187, 15, 158, 31]. Bridging the gap between theory and practice, therefore, can accelerate progress in the field, providing new opportunities to formulate clearer research objectives and improve the evaluation methodologies [16, 158, 69, 24].

As an attempt to provide an epistemologically grounded characterisation for Explanation-based NLI, this chapter aims to bridge the gap in the notion of a *scientific explanation* [131, 132], studying it as both a *formal object* and as a *linguistic expression*.

To this end, the chapter is divided in two main sections. The first part represents a systematic survey of the modern discussion in Philosophy of Science, shedding light on the nature and function of explanatory arguments and their constituting elements [55, 91]. Following the survey, the second part of the chapter presents a corpus analysis aimed at qualifying sentence-level *explanatory patterns* in corpora of natural language explanations, focusing on datasets used to build and evaluate explanation-based inference models in the scientific domain [179, 70].

Overall, the chapter presents the following main conclusions:

1. **Explanations cannot be exclusively characterised in terms of *inductive* or *deductive* arguments.** Specifically, the main function of an explanation is not of *predicting* or *deducing* the event to be explained (*explanandum*) [56], but the one of showing how the explanandum fits into a *broader underlying regularity*. This process is known as *unification*, and it is responsible for the creation of *explanatory patterns* that can account for a large set of phenomena [47, 90].

2. **An explanation must cite part of the causal history of the explanandum**, fitting the event to be explained into a *causal nexus* [132]. There are two possible ways of constructing causal explanations: (1) an explanation can be *etiological* – i.e., the explanandum is explained by revealing part of its causes – or (2)

constitutive – i.e., the explanation describes the underlying mechanism giving rise to the explanandum. This is confirmed by the corpus analysis, which reveals that the majority of natural language explanations, indeed, contain references to mechanisms and/or direct causal interactions between entities [70].

3. **While explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms.** In particular, additional knowledge categories such as *definitions*, *properties* and *taxonomic relations* seem to play an equally important role in building an explanatory argument. This can be attributed to both *pragmatic aspects* of natural language explanations as well as inference requirements associated to *unification*.

4. **Patterns of unification naturally emerge in corpora of explanations.** Even if not intentionally modelled, *unification* seems to be an emergent property of corpora of natural language explanations [179]. The corpus analysis, in fact, reveals that the distribution of certain explanatory sentences is connected to the notion of *unification power* and that it is possible to draw a parallel between inference patterns emerging in natural language explanations and formal accounts of explanatory unification [91].

5. **Unification is realised through a process of abstraction.** Specifically, abstraction represents the fundamental inference substrate supporting unification in natural language, connecting concrete instances in the explanandum to high-level concepts in central explanatory sentences. This process, realised through specific linguistic elements such as definitions and taxonomic relations, is a fundamental part of natural language explanations, and represents what allows subsuming the event to be explained under high-level patterns and unifying regularities.

We conclude by suggesting how the presented findings can open new lines of research for Explanation-based NLI systems, informing the way the community should approach model creation and the design of evaluation methodologies for natural language explanations.

The chapter contributes to addressing a fundamental gap between classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts. This characterisation can support a more principled design of systems that can better interpret and generate natural language explanations. To the best of our knowledge, while previous work on natural language explanations have performed

Account	Explanans	Relation
Epistemic		
Deductive-Nomological [55]	Initial conditions + at least a universal law of nature	The *explanandum* is logically deduced from the *explanans*
Inductive-Statistical [56]	Initial conditions + at least a statistical law	The *explanans* make the *explanandum* highly probable
Unificationist [91]	A theory T + a class of phenomena P including the *explanandum*	Shows how a class of phenomena P can be derived from a theory T through the instantiation of an argument pattern
Ontic		
Statistical-Relevance [133]	A set of statistically relevant facts	the *explanans* increase the probability of the *explanandum*
Causal-Mechanical [132]	A set of relevant causal processes and interactions	The *explanans* are part of the causal history of the *explanandum*; the *explanans* are part of the mechanism underlying the *explanandum*

Table 2.1: The main modern accounts of scientific explanation in Philosophy of Science.

quantitative and qualitative studies in terms of knowledge reuse and inference categories [67, 73], this study is the first to explore the relation between linguistic aspects of explanations and formal accounts in Philosophy of Science [178].

2.2 The Epistemological Perspective

The ultimate goal of science goes far beyond the pure prediction of the natural world. Science is constantly seeking a deeper understanding of observable phenomena and recurring patterns in nature by means of scientific theories and explanations. Most philosophers define an explanation as an answer to a *"why"* question, aiming at identifying and describing the reason behind the occurrence and manifestation of particular events [132]. However, although the explanatory role of science is universally acknowledged, a formal definition of what constitutes and characterise a scientific explanation remains a complex subject. This is attested by the long history of the discussion in Philosophy of Science, which goes back at least to Ancient Greece [53]. Nevertheless, relatively recent attempts at delivering a rigorous account of scientific explanation have produced a set of quasi-formal models that clarify to some extent the nature of the concept [178, 131, 169].

The modern view of scientific explanation has its root in the work of Carl Gustav Hempel and Paul Oppenheim, *"Studies in the Logic of Explanation"* [55], whose publication in 1948 raised a heated debate in the Philosophy of Science community [178]. This section will survey the main accounts resulting from this debate with the aim of summarising and revisiting the main properties of a scientific explanation. In particular, the goal of the survey is to identify the principal constraints that these models impose on *explanatory arguments*, highlighting their essential features.

In general, an explanation can be described as an argument composed of two main elements [132]:

1. The *Explanandum*: the fact representing the observation or event to be explained.
2. The *Explanans*: the set of facts that are invoked and assembled to produce the explanation.

The aim of a formal account of a scientific explanation is to define an *"objective relationship"* that connects the explanandum to the explanans [132], imposing constraints on the class of possible arguments that constitute a valid explanation. Existing accounts, therefore, can be classified according to the nature of the relationship between explanans and explanandum (Table 2.1). Specifically, this survey will focus on accounts falling under two main conceptions [132]:

- *Epistemic:* The explanation is an *argument* showing how the explanandum *can be derived* from the explanans. There is a relation of *logical necessity* between the explanatory statements and the event to be explained.
- *Ontic:* The explanation relates the explanandum to *antecedent conditions* by means of general laws, *fitting* the explanandum into a *discernible pattern*.

2.2.1 Explanation as an Argument

Deductive-Inductive Arguments

The *Deductive-Nomological (DN)* model proposed by Hempel [55] is considered the first modern attempt to formally characterise the concept of scientific explanation. the DN account defines an explanation as an argument, connecting explanans and explanandum by means of *logical deduction* [178]. Specifically, the explanans constitute the premises of a deductive argument while the explanandum represents its logical conclusion. The general structure of the DN model can be schematised as follows:

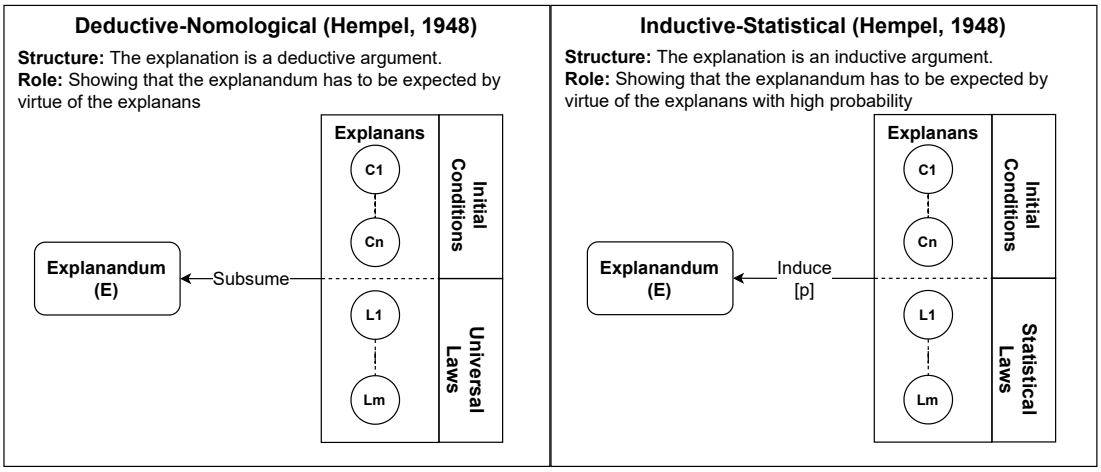


Figure 2.1: Schematic representation of the Deductive-Inductive account of scientific explanation.

$$\begin{array}{ll} C_I, C_2 \ldots, C_k & \text{Initial Conditions} \\ \hline L_I, L_2 \ldots, L_r & \text{Universal Laws of Nature} \\ E & \text{Explanandum} \end{array}$$

In this model, the explanans are constituted by a set of initial conditions, $C_1, C_2 \ldots, C_k$, plus at least a universal law of nature, $L_1, L_2 \ldots, L_r$ (with $r > 0$). According to Hempel, in order to represent a valid scientific argument, an explanation must include only explanans that are empirically testable. At the same time, the universal law must be a statement describing a *universal* regularity, while the initial conditions represent particular facts or phenomena that are concurrently observable with the explanandum. Here is a concrete example of a scientific explanation under the DN account [56]:

- C_1: The (cool) sample of mercury was placed in hot water;
- C_2: Mercury is a metal;
- L_1: All metals expand when heated;
- E: The sample of mercury expanded.

To complete the DN account with a theory of statistical explanation, Hempel introduced the *Inductive-Statistical (IS)* model [56]. According to the IS account, an explanation must show that the explanandum was to be expected with *high probability* given the explanans. Specifically, an explanation under the IS account has the same structure of the DN account, replacing the universal laws with statistical laws. In order

for a statistical explanation to be appropriate, the explanandum must be induced from statistical laws and initial conditions with probability close to 1.

The Deductive-Inductive view proposed by Hempel emphasises the *predictive power* of explanations. Given a universal/statistical law and a set of initial conditions, it is possible to establish whether or not a particular phenomenon will occur in the future. According to Hempel, in fact, not only predictive power is a fundamental property of an explanation, explanations and predictions share exactly the *same logical structure*. Specifically, the only difference between explanatory and predictive arguments is when they are formulated or requested: explanations are generally required for past phenomena, while predictions for events that have yet to occur.

This feature of the Deductive-Inductive account is known as the *symmetry thesis* [56] which has been largely criticised by other philosophers in the field [132, 91, 178, 169]. The symmetry thesis, in fact, leads to well-known objections and criticisms of Hempel's account. Consider the following example [91, 131]:

- C_1: The elevation of the sun in the sky is x;
- C_2: The height of the flagpole is y;
- L_1: Laws of physics concerning the propagation of light;
- L_2: Geometric laws;
- E: The length of the shadow is z.

While the example above represents a reasonable explanatory argument, the DN account does not impose any constraint that prevents the interchanging of the explanandum with some of the initial conditions:

- C_1: The elevation of the sun in the sky is x;
- C_2: The length of the shadow is z;
- L_1: Laws of physics concerning the propagation of light;
- L_2: Geometric laws;
- E: The height of the flagpole is y.

The DN model and its symmetry property, in particular, allows for the construction of explanatory arguments that contain inverted causal relations between its elements. This

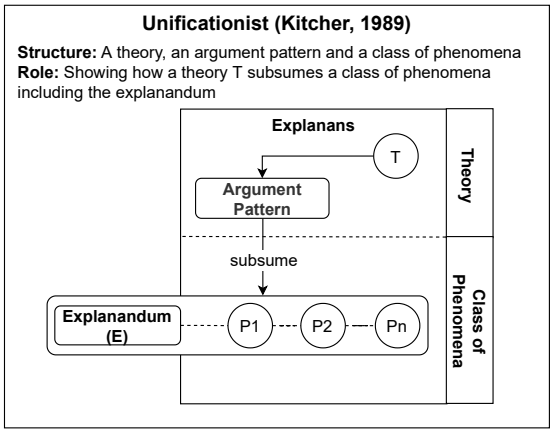


Figure 2.2: A schematic representation of the Unificationist account of scientific explanation.

counterexample shows that prediction and explanation *must have a different logical structure* and treated as different types of arguments. Although predictive power is a necessary property of an adequate explanation, it is not sufficient. Explanations, in fact, are inherently *asymmetric*, a property that cannot be described by means of deductive-inductive arguments alone.

In Hempel's account, moreover, there is a further property of explanation that has been subject to criticisms by subsequent philosophers, that is the notion of *explanatory relevance*. Consider the following counter-example from Salmon [132]:

- C_1: John Jones is a male;
- C_2: John Jones has been taking birth control pills regularly;
- L_1: Males who take birth control pills regularly fail to get pregnant;
- E: John Jones fails to get pregnant.

Although the argument is formally correct, it contains statements that are explanatorily irrelevant to E. Specifically, the fact that *John Jones has taken birth control pills* should not be cited in an explanation for *John Jones fails to get pregnant*. In this particular example only C_1 is relevant to E, and only C_1 should figure into an explanation for E. Specifically, the universality and high probability requirements of the DN and IS model constrain the explanation to include all the explanatory relevant premises but not to exclude irrelevant facts [132].

Explanatory Unification and Argument Patterns

The Unificationist account of scientific explanation was proposed by Friedman [47] and subsequently refined by Kitcher [91, 90] in order to overcome the criticisms, including relevance and asymmetry, raised by the Deductive-Inductive account.

According to the Unificationist model, an explanation cannot be uniquely described in terms of deductive or inductive arguments. To properly characterise an explanation, in fact, it is necessary to consider its main function of fitting the explanandum into a *broader unifying pattern*. Specifically, an explanation is an argument whose role is to connect a set of *apparently unrelated phenomena*, showing that they can be subsumed under a common underlying regularity. The concept of explanatory unification is directly related to the goal of Science of understanding nature by reducing the number of disconnected phenomena and provide an ordered and clear picture of the world [138].

Figure 2.2 shows a schematic representation of the Unificationist account. Given a scientific theory T and a class of phenomena P including the explanandum E, an explanation is an argument that allows deriving all the phenomena in P from T. In this case, we say that T *unifies* the explanandum E with the other phenomena in P. According to Kitcher, a scientific explanation accomplishes unification by deriving descriptions of many phenomena through the same patterns of derivation [91]. Specifically, a theory defines an *argument pattern* which can be occasionally instantiated to explain particular phenomena or observations.

An argument pattern is a sequence of *schematic sentences* organised in premises and conclusions. In particular, a schematic sentence can be described as a template obtained by replacing some non-logical expressions in a sentence with *variables* or *dummy letters* [90, 91, 176]. For instance, according to Kitcher, from the statement *"Organisms homozygous for the sickling allele develop sickle-cell anemia"* it is possible to generate schematic sentences at different levels of abstraction: *"Organisms homozygous for A develop P"* and *"For all x, if x is O and A then x is P"*. An argument pattern can be instantiated by specifying a set of *filling instructions* for replacing the variables of the schematic sentences together with rules of inference for the derivation. Following the previous example, a possible filling instruction for the schematic sentence *"Organisms homozygous for A develop P"* might specify that A must be substituted by the name of an allele and P by some phenotypic trait [90, 91]. Different theories can induce different argument patterns whose structure is not defined a-priori as in the case of Hempel's account. However, once a theory is accepted, the same argument pattern can be instantiated to explain a large variety of phenomena depending on the unification

power of the theory.

The history of science is full of theories and explanations performing unification, and the advancement of science itself can be seen as a process of growing unification [47]. A famous example is provided by Newton's law of universal gravitation, which unifies the motion of celestial bodies and falling objects on Earth showing that they are all manifestation of the same underlying physical law. Specifically, from the unificationist point of view, Newton's law of universal gravitation defines an argument pattern whose filling instructions apply to all entities with mass.

The Unificationist account provides a set of criteria to identify the *"best explanation"* among competing theories [90, 91]:

1. *Unification power*: Given a set of phenomena P and a theory T. the larger is the cardinality of P - i.e. the number of phenomena that are unified by T, the greater is the explanatory power of T.

2. *Simplicity*: Given two theories T and T_1 able to unify the same set of phenomena P, the theory that makes use of a lower number of premises in its argument patterns is the one with the greatest explanatory power.

These selection criteria play a fundamental role in the Unificationist account since, according to Kitcher, only the best explanation available at a given point in time should be considered as the valid one [90]. For example, to explain the motion of celestial bodies by means of gravity, one must consider Einstein's theory of relativity as the valid explanation, as it allows subsuming a broader set of phenomena compared to Newton's law of universal gravitation.

The simplicity criterion prevents the explanation to include irrelevant premises as in the case of the control pill example analysed under the Deductive-Inductive account since, under the same unification power, an explanation containing less premises will be preferred over a more complex explanation introducing unnecessary statements. Similarly, the problem of asymmetry can be solved considering the unification power criterion. Specifically, argument patterns containing inverted causal relations will generally allow for the derivation of fewer phenomena. According to Kitcher, in fact, causality is an emergent property of unification: *"to explain is to fit the phenomena into a unified picture insofar as we can. What emerges in the limit of this process is nothing less than the causal structure of the world"* [91].

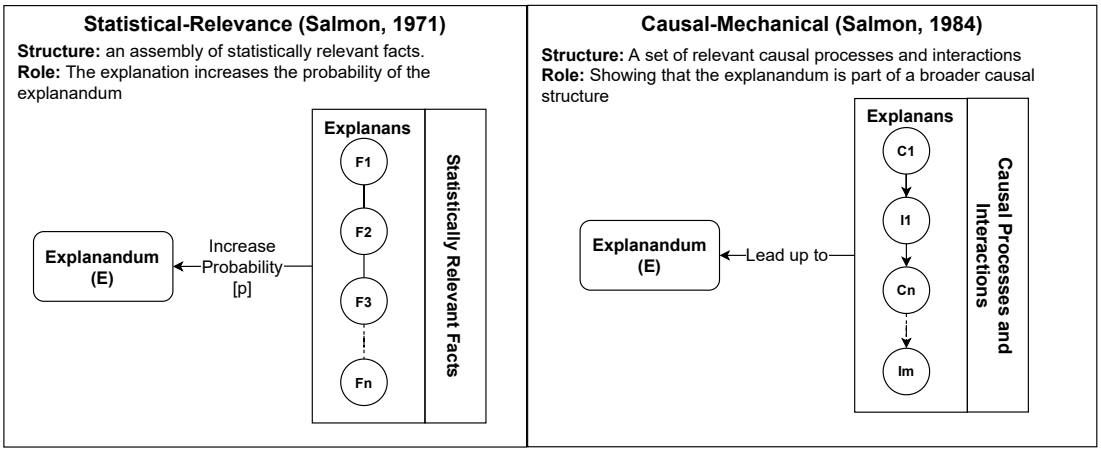


Figure 2.3: Schematic representation of accounts falling under the *ontic* conception.

2.2.2 Fitting the Explanandum into a Discernible Pattern

Statistical-Relevance

Motivated by the problem of relevance in the Deductive-Inductive account, Wesley Salmon elaborated a statistical account of explanation known as *Statistical Relevance (SR)* [133]. Differently from the Deductive-Inductive account, the SR model is not concerned with the general structure and organisation of the explanatory argument, but attempts to characterise a scientific explanation in terms of the intrinsic relation between each explanatory statement and the explanandum.

In general, given a population A, a factor C and some event B, we say that C is *statistically relevant* to the occurrence of B if and only if

$$P(B|A.C) \neq P(B|A) \vee P(B|A.C) \neq P(B|A.\neg C) \tag{2.1}$$

In other words, a given factor C is statistically relevant to an event B if its occurrence changes the conditional probability of B to occur [133, 132]. According to the SR account, the explanatory relevance of a fact has to be defined in terms of its statistical relevance. Specifically, an explanation is an *assembly of statistically relevant facts* that increase the probability of the explanandum.

Consider the birth control pills example analysed under the IS account [132]:

- C_1: John Jones is a male;
- C_2: John Jones has been taking birth control pills regularly;
- E: John Jones fails to get pregnant.

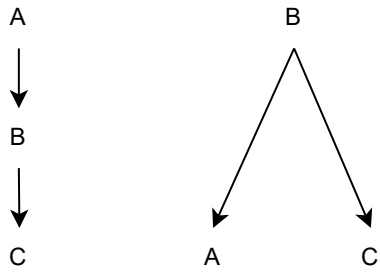


Figure 2.4: Causal relationships are underdetermined by statistical relevance relationships. In this example, in particular, it is not possible to discriminate between the depicted causal structures using a statistical relevance analysis. In both cases, in fact, A is statistically relevant to C; a factor that can lead, in the situation depicted on the right, to a SR explanation based on the relation between A and C induced by the common cause B.

According to Salmon [133, 132], given a population T, we can perform a statistical analysis to verify whether C_1 and C_2 are relevant to E:

$$P(pregnant|T.male) = P(pregnant|T.male.pills) \tag{2.2}$$

$$P(pregnant|T.pills) \neq P(pregnant|T.pills.male) \tag{2.3}$$

Notice that in (2.2), given the fact that a generic $x \in T$ is a male ($T.male$), the action of taking birth control pills ($T.male.pills$) has no affect on the probability that x is pregnant. Conversely, in (2.3), the probability that a generic member of the population x is pregnant, given the action of taking pills ($T.pills$), decreases to zero if we know that x is a male ($T.pills.male$). Therefore, the statistical relevance analysis leads to the conclusion that *"among males, taking birth control pills is explanatorily irrelevant to pregnancy, while being male is relevant"* [132].

Although statistical relevance seemed to provide a formal way to shield explanation from irrelevance, Salmon subsequently realised that the SR model is not sufficient to elaborate an adequate account of scientific explanation [132, 130]. It is nowadays clear, in fact, that certain causal structures are greatly underdetermined by statistical relevance [119, 120]. Specifically, different causal structures can be described by the same statistical relevance relationships among their elements, making it impossible to discriminate direct causal links by means of a statistical relevance analysis alone (Figure 2.4).

According to Salmon, *"the statistical relationships specified in the SR model constitute the statistical basis for a bona-fide scientific explanation, but this basis must be*

supplemented by certain causal factors in order to constitute a satisfactory scientific explanation" [132]. The failed attempt to characterise a scientific explanation uniquely in terms of statistical elements demonstrated, as in the case of Hempel's account, the intrinsic difference between prediction and explanation. The latter, in fact, cannot be derived by pure statistical observations and seems to require conjectures and hypotheses about hidden structures, such as the one induced by causal relations and interactions.

Causes and Mechanisms

Following the observation that the SR model is not sufficient for characterising a scientific explanation, Salmon formulated a new account known as the Causal-Mechanical (CM) model [132], in which the role of an explanation is to show how the explanandum fits into the *causal structure of the world*. Specifically, a valid scientific explanation cannot be limited to statistical relevance and must *cite part of the causal history* leading up to the explanandum.

To formalise the CM account, Salmon attempted to define a theory of causality based on the concepts of *causal processes* and *interactions* [130]. Consider the following example from Woodward [176]: *"a cue ball, set in motion by the impact of a cue stick, strikes a stationary 8 ball with the result that the 8 ball is put in motion and the cue ball changes direction"*. Here, the cue ball, the cue stick and the 8 ball are *causal processes* while the collisions between the objects are *causal interactions*. According to the CM model, the motion of the 8 ball has to be explained in terms of the causal processes and interactions belonging to its causal history. Therefore, a generic event X is explanatorily relevant to the explanandum E if and only if the following conditions apply [132]:

1. X is statistically relevant to E
2. X and E are part of different causal processes
3. There exists a sequence of causal processes and interactions between X and E leading up to E

Salmon identifies two major ways of constructing causal explanations for some event E. An explanation can be either *etiological* – i.e. E is explained by revealing part of its causes – or *constitutive* – i.e. the explanation of E describes the underlying mechanism giving rise to E. A mechanism, in particular, is often described as an organised set of entities and activities, whose interaction is responsible for the emergence of a phenomenon [175, 25, 27]. For example, it is possible to formulate an etiological

Type of implied question	Type of contrast case	Type of cause
"Why X rather than not X?"	Non occurrence of effect	Sum of necessary conditions
"Why X rather than the default value for X?"	The normal case	Abnormal condition
"Why X rather than Y?"	Noncommon effect	Differentiating factor
"Why X rather than what ought to be the case?"	Prescribed or statutory case	Moral or legal fault
"Why X rather than the ideal value for X?"	Ideal case	Design fault or bug

Table 2.2: Different causal questions and attributions with different implied contrast cases as defined in [59].

explanation of a certain disease by appealing to a particular virus, or we can provide a constitutive explanation describing the underlying mechanisms by which the virus causes the disease.

The foremost merit of the CM account is to exhibit the profound connection between causality, mechanisms, and explanation, highlighting how most of the fundamental characteristics of a scientific explanation derive from its causal nature. Moreover, the differentiation between etiological and constitutive explanation had a significant impact on several scientific fields. Discovering mechanistic explanations, in fact, is nowadays acknowledged as the ultimate goal of many scientific disciplines such as biology and natural sciences [28, 136, 25, 9].

The CM model is still subject to a number of criticisms concerning the concepts of causal processes and interactions, which has led subsequent philosophers to propose new theories of causality [99, 176, 60, 177]. However, the causal nature of scientific explanations is largely accepted, with much of the contemporary discussion focusing on philosophical and metaphysical aspects concerning causes and effects [119].

An additional criticism is related to the inherent incompleteness of causal explanations [57, 29]. Since the causes of some event can be traced back indefinitely, causal explanations must show only part of the causal history of the explanandum. This implies that not all the causes of an event can be included in an explanation. In Salmon's account, however, it is not clear what are the criteria that guide the inclusion of relevant causes and the exclusion of others. Subsequent philosophers claimed that the problem of relevance is context-dependent and that it can be only addressed by looking at explanations from a pragmatic perspective [162]. All why questions, in fact, seem to be *contrastive* in nature [105, 112]. Specifically, once a causal model is known, the

explanans selected for a particular explanation depend on the specific why question, including only those causes that *make the difference* between the occurrence of the explanandum and some *contrast case* implied by the question [113, 59] (Table 2.2).

2.2.3 Summary

This section presented a detailed overview of the main modern accounts of scientific explanation, discussing their properties and limitations.

Despite a number of open questions remain in the Philosophy of Science community [131, 178], it is possible to draw the following conclusions:

1. **Explanations and predictions have a different structure.** Any attempt to characterise a scientific explanation uniquely in terms of predictive elements has encountered fundamental issues from both an epistemic and an ontic perspective. An explanation, in fact, cannot be entirely characterised in terms of *deductive-inductive arguments* or *statistical relevance* relationships. This is because predictive power, despite being a necessary property of a scientific explanation, is not a sufficient one.

2. **Explanatory arguments create unification.** From an epistemic perspective, the main function of an explanatory argument is to fit the explanandum into a *broader unifying pattern*. Specifically, an explanation must connect a class of *apparently unrelated phenomena*, showing that they can be subsumed under a common underlying regularity. Form a linguistic point of view, the unifying power of explanations produces *argument patterns*, whose instantiation can be used to explain a large variety of phenomena through the same patterns of derivation.

3. **Explanations possess an intrinsic causal-mechanistic nature.** From an ontic perspective, a scientific explanation must cite part of the causal history of the explanandum, fitting the event to be explained into a *causal nexus*. There are two possible ways of constructing causal explanations: (1) an explanation can be *etiological* – i.e., the explanandum is explained by revealing part of its causes – or (2) *constitutive* – i.e., the explanation describes the underlying mechanism giving rise to the explanandum.

Philosophers tend to agree that the causal and unificationist accounts are complementary to each other, advocating for a *"peaceful coexistence"* and a pluralistic view of scientific explanation [131, 178, 145, 8, 50]. Unification, in fact, seems to be an

essential property of causal explanations since many physical processes are the result of the same underlying causal mechanisms [130, 131, 8]. At the same time, the unifying power of constitutive explanations derives from the existence of mechanisms that have a common higher-level structure, despite differences in the specific entities composing them [50].

Moreover, the unificationist account seems to be connected with theories of explanation and understanding in cognitive science, which emphasise the relationship between the process of searching for broader regularities and patterns to the way humans construct explanations in everyday life through abductive reasoning, abstraction, and analogies [110, 109, 79, 149].

2.3 The Linguistic Perspective

The previous section focused on the notion of a scientific explanation from a quasi-formal (categorical) perspective, reviewing the main epistemological accounts attempting to characterise the space of valid explanatory arguments. Following this survey, this section assumes a linguistic perspective, investigating how the main features of the accepted accounts manifest in *natural language*.

To this end, we present a systematic analysis of corpora of scientific explanations in natural language adopting a mixture of qualitative and quantitative methodologies to investigate the emergence of *explanatory patterns* at both *sentence* and *inter-sentence* level, relating them to the *Causal-Mechanical* [130] and *Unificationist* account [90, 91]. Specifically, we hypothesise that it is possible to map linguistic aspects emerging in natural language explanations to the discussed models of scientific explanation. At the same time, we observe that some linguistic and pragmatic elements in natural language explanations are not considered by the epistemological accounts, and therefore expect the corpus analysis to provide complementary insights on the nature of explanations as manifested in natural language. Bridging the gap between these two domains aims to provide a linguistic-epistemological grounding for the construction of Explanation-based NLI models.

The presented analysis focuses on two distinct corpora of explanations; the *Biology Why Corpus* [2] [70], a dataset of biology why-questions with one or more explanatory passages identified in an undergraduate textbook, and the *WorldTree Corpus* [3]

[2] https://allenai.org/data/biology-how-why-corpus
[3] http://cognitiveai.org/explanationbank/

Feature	Why Corpus	WorldTree
Size	193	2206
Domain	Biology	Science exams
Type	Scientific	Scientific - Commonsense
Annotation	Textbooks	Manually curated
Structured	No	Yes
Reuse	No	Yes

Table 2.3: Main features of the analysed explanations corpora.

[179], a corpus of science exams questions curated with natural language explanations supporting the correct answers.

The main features of the selected corpora are summarised in Table 2.3. As shown in the table, the corpora have complementary characteristics. The explanations included in the *Biology Why Corpus* are specific to a scientific domain (biology in this case), while the *WorldTree Corpus* expresses a more diverse set of topics, including physics, biology, and geology. Since the explanatory passages from the *Biology Why Corpus* are extracted from textbooks, the explanations tend to be more technical and unstructured. On the other hand, the explanations in *WorldTree* are manually curated and represented in a semi-structured format (aiming more closely at inference automation), often integrating scientific sentences with commonsense knowledge. Moreover, the individual explanatory sentences in *WorldTree* are reused across different science questions when possible, facilitating a quantitative study on knowledge use and the emergence of sentence-level explanatory patterns [73].

By leveraging the complementary characteristics of the selected corpora and relating the corpus analysis to the discussed accounts of scientific explanation, we aim at investigating the following research questions:

1. **RQ1:** What kinds of explanatory sentences occur in natural language explanations?

2. **RQ2:** How do explanatory patterns emerge in natural language explanations?

We adopt the *Biology Why Corpus* and *WorldTree* to investigate **RQ1**, while *WorldTree* is considered for **RQ2** due to its size and structure.

Explanandum	Explanans	Knowledge Category
It is important for blood transfusions to not occur between individuals with different blood types	Certain bacteria normally present in the body have epitopes very similar to the A and B carbohydrates	Analogy, Comparison
It is important for blood transfusions to not occur between individuals with different blood types	By responding to the bacterial epitope similar to the B carbohydrate, a person with type A blood makes antibodies that will react with the type B carbohydrate	Process, Mechanism
It is important for blood transfusions to not occur between individuals with different blood types	Matching compatible blood groups is critical for safe blood transfusions	Requirement, Constraint
Inbreeding does not cause evolution directly	The Hardy-Weinberg is a principle that describes a hypothetical population that is not evolving	Definition
Inbreeding does not cause evolution directly	The gene pool is modified if mutations alter alleles or if entire genes are deleted or duplicated	Conditional, If-then
Inbreeding does not cause evolution directly	Both inbreeding and genetic drift can cause a loss of genetic variation	Causal Interaction
Inbreeding does not cause evolution directly	The allele and genotype frequencies often do change over time	Property, Attribute
Steroids can easily pass through cell membranes	These complexes of a lipid-soluble hormone and its receptor act in the nucleus to regulate transcription of specific genes	Function, Roles
Chromatin is important in meiosis	For example, the nuclei of human somatic cells (all body cells except the reproductive cells) each contain 46 chromosomes	Instances, Examples
It is important for polypeptides to be able to greatly vary in amino acid sequence	Recall that most enzymes are proteins	Taxonomic, Meronymic
Two traits that are more than 50cM away from each other are inherited randomly relative to each other	The observed frequency of recombination in crosses involving two such genes can have a maximum value of 50%	Statistical Relations

Table 2.4: Explanation sentences in the Biology Why Corpus.

2.3.1 Biology Why Questions

To study and investigate the emergence of sentence-level explanatory patterns in biological explanations we performed a systematic annotation of the explanatory passages included in the *Biology Why Corpus* [70]. To this end, we identified a set of 11 recurring knowledge categories, annotating a sample of 50 explanations extracted from the corpus. Examples of annotated explanation sentences and their respective knowledge categories are included in Table 2.4. The complete set of annotated explanations adopted in the corpus analysis is available online[4].

[4] https://github.com/ai-systems/scientific_explanations_analysis

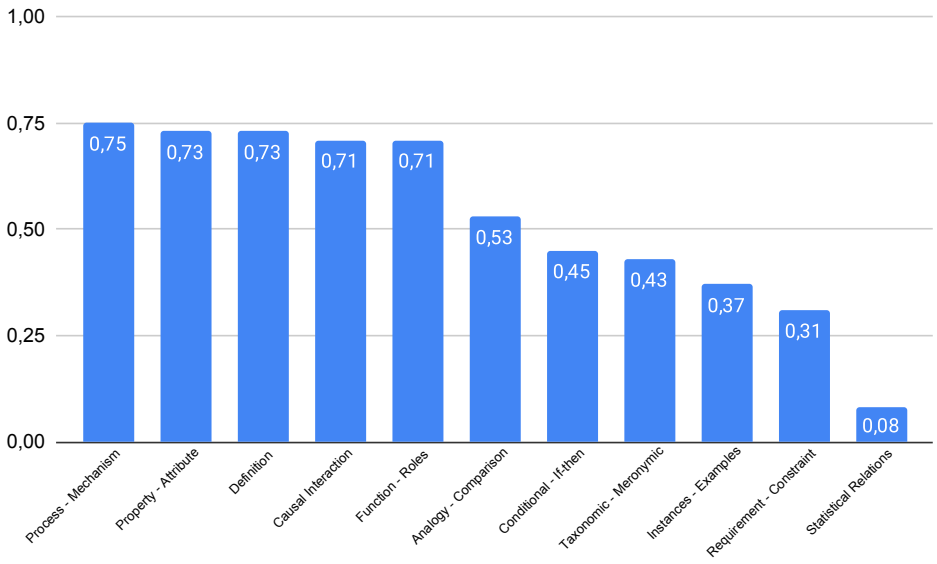


Figure 2.5: Recurring knowledge in biological explanations.

Recurring Explanatory Sentences

Figure 2.5 reports the frequencies of each knowledge category in the annotated why-questions. Specifically, we consider each knowledge category as a binary variable (1 if the knowledge category appears in an answer to a why question, 0 otherwise) and compute a binomial distribution for each type.

The corpus analysis reveals that the majority of the why questions (75%) are answered through the direct description of *processes* and *mechanisms*. As expected, this result confirms the crucial role of *constitutive explanations* as defined in the Causal-Mechanical (CM) account [132]. The importance of causality is confirmed by the frequency of sentences describing direct *causal interactions* between entities (71%), which demonstrates the interplay between *constitutive* and *etiological* explanations. Moreover, the analysis suggests that a large part of the explanations (71%) include sentences describing *functions* and *roles*. The relation between the notion of function and mechanisms is reported in many constitutive accounts of explanation [25], and is typically understood as a mean of describing and situating some lower-level part within a higher-level mechanism [26].

The corpus analysis suggests that natural language explanations are not limited to causes and mechanisms and tend to include additional types of knowledge not explicitly discussed in the epistemological accounts. Specifically, the graph reveals that *definitions* and sentences about *attributes* and *properties* play an equally important role

Explanandum
Two sticks getting warm when rubbed together is an example of a force producing heat.
Explanans
(1) A stick is a kind of object; (2) To rub together means to move against; (3) Friction is a kind of force; (4) Friction occurs when two object's surfaces move against each other; (5) Friction causes the temperature of an object to increase.

Table 2.5: Example of a curated explanation in WorldTree.

in the explanations (both occurring in 73% of the why questions). We attribute this result to *pragmatic aspects* and inference requirements associated to the *unification* process. Definitions, for instance, might serve both as a way to introduce missing context and background knowledge in natural language discourse and, in parallel, as a mechanism for *abstraction*, relating specific terms to high-level conceptual categories [139, 140, 144].

The role of abstraction in the explanations is supported by the presence of *analogies* and *comparison* between entities (53%), as well as sentences describing *taxonomic* or *meronymic* relations (43%). These characteristics suggest the presence of explanatory arguments performing unification through an abstractive inference process, whose function is to identify common abstract features between concrete instances in the explanandum [90]. The role of abstraction will be explored in details in the next section.

Finally, the corpus analysis reveals a low frequency of sentences describing *statistical relevance* relationships and *probabilities* (8%). These results reinforce the fundamental difference between explanatory and predictive arguments identified and discussed in the philosophical accounts [178, 169].

2.3.2 Science Questions

This section presents a corpus analysis on WorldTree [179] aimed at investigating the emergence of explanatory patterns and unification, relating them to epistemological aspects of scientific explanations. Table 2.5 shows an archetypal example of explanation in WorldTree. Here, the explanandum is represented by a statement derived from a science question and its correct answer, while the explanans are an assembly of sentences retrieved from a background knowledge base.

The corpus categorises the core explanans according to different explanatory roles:

- *Central:* Sentences explaining the central concepts that the question is testing.

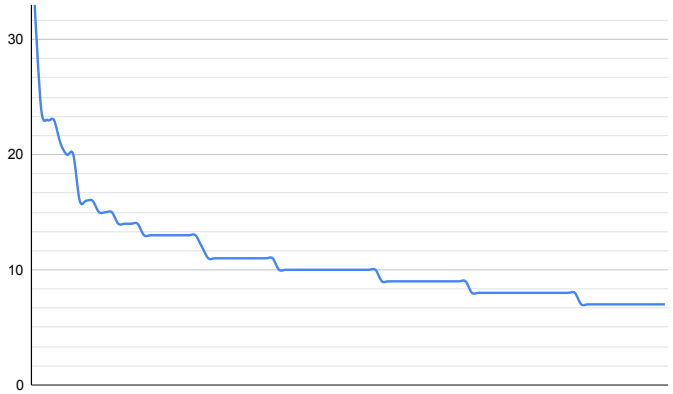


Figure 2.6: Distribution and reuse of central explanatory sentences in WorldTree. The y axis represents the number of times a central sentence appears in the explanations included in the corpus, while the points on the x axis represent each individual central sentence in the corpus.

- *Grounding:* Sentences linking generic terms in a central sentence with specific instances of those terms in the question.

Some explanatory sentences in WorldTree can be categorised according to additional roles that are not strictly required for the inference (i.e., *Background* and *Lexical Glue* [72]) and that, for the purpose of investigating the nature of explanatory patterns, will not be considered in the corpus analysis.

Distribution and Reuse of Explanatory Sentences

The first analysis concentrates on the distribution and reuse of *central* explanatory sentences in the corpus. The quantitative results of this analysis are presented in Figure 2.6 and 2.7, while a set of qualitative examples are reported in Table 2.6.

The graph in Figure 2.6 describes the distribution of individual sentences annotated as central explanatory facts across different explanations. Specifically, the y-axis represents the number of times a specific sentence is used as a central explanation for a specific science question. The trend in the graph reveals that the occurrence of central explanatory sentences tends to follow a long tail distribution, with a small set of sentences frequently reused across different explanations. This trend suggests that a subset of sentences results particularly useful to construct explanations for many science questions, constituting a first indication that some central sentence might possess a greater *explanatory power* and induce certain *patterns of unification*.

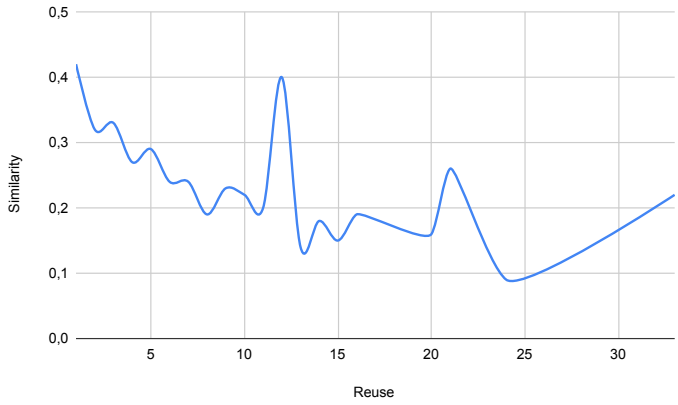


Figure 2.7: Similarity between central sentences and questions vs frequency of reuse of the central sentences. The y axis represents the average similarity (ranging between 0 and 1) between a central sentence and the questions it explains, while the x axis represents the number of times the central sentence appears in the explanations included in the corpus.

To further investigate this aspect, Figure 2.7 correlates the frequencies of central explanatory sentences in the corpus (x axis) with the average similarity between the same sentences and the questions they explain (y axis). To perform the analysis, the similarity values are computed adopting BM25 and cosine distance between each question and its explanation sentences [129]. From a unificationist point of view, we expect to find an inverse correlation between the frequency of reuse of a central sentence and its similarity with the explanandum. Specifically, we assume that the lower the similarity, the higher the probability that a central sentence describes abstract laws and high level regularities, and that, therefore, it is able to *unify* a larger set of phenomena. Under this assuptions and considering naturally occurring variability in the dataset, the trend in Figure 2.7 confirms the expectation, showing that the most reused central sentences are also the one that explain clusters of less similar questions. In particular, the graph reinforces the hypothesis that the reuse value of a central sentence in the corpus is indeed connected with its *unification power*.

The concrete examples in Table 2.6 further support this hypothesis. Specifically, the table shows that it is possible to draw a parallel between the distribution of central sentences in the corpus and the notion of *argument patterns* in the Unificationist account [90]. It is possible to notice, in fact, that the most occurring central sentences tend to describe high-level processes and regularities, typically mentioning abstract concepts and entities (e.g., *living things*, *object*, *substance*, *material*). In particular, the examples

Central Explanatory Sentence	Occurrence
Boiling;evaporation means matter; a substance changes from a liquid into a gas by increasing heat energy	33
An adaptation; an ability has a positive impact on an animal's; living thing's survival; health; ability to reproduce	24
Photosynthesis means producers; green plants convert from carbon dioxide and water and solar energy into carbohydrates and food and oxygen for themselves	23
Inheriting is when an inherited characteristic is copied; is passed from parent to offspring by genetics; DNA	23
Melting means matter; a substance changes from a solid into a liquid by increasing heat energy	21
If an object is made of a material then that object has the properties of that material	20
Photosynthesis is a source of; makes food; energy for the plant by converting carbon dioxide, water, and sunlight into carbohydrates	20
Water is in the solid state , called ice , for temperatures between 0; -459; -273 and 273; 32; 0 K; F; C	16
Decomposition is when a decomposer breaks down dead organisms 16 an animal; living thing requires nutrients for survival	16
Objects are made of materials; substances; matter	15
Chemical reactions cause new substances; different substances to form	15

Table 2.6: Most reused central explanatory sentences in WorldTree.

suggest that reoccurring central explanatory facts might act as *schematic sentences* of an *argument pattern*, with abstract entities representing the linguistic counterpart of *variables* and *filling instructions* used to specify and constraining the space of possible instantiations.

Abstraction and Patterns of Unification

To further explore the parallel between natural language explanations and the Unificationist account, we focus on recurring inference chains between *grounding* and *central* sentences. Specifically, we aim at investigating whether it is possible to map inference patterns in WorldTree to the process of instantiating *schematic sentences* for unification. To this end, we automatically build a linkage between grounding and central sentences in the corpus using the support of lexical overlaps.

Table 2.7 reports the most recurring linguistic categories of grounding-central chains, which provide and indication of the high-level process through which explanatory patterns emerge in natural language. Overall, we found a clear evidence of inference patterns related to the *instantiation* of central explanatory sentences. Specifically, the

Grounding	Grounding	Occurrence
_ is a kind of _ (Taxonomic)	_ is a kind of _ (Taxonomic)	524
_ is a kind of _ (Taxonomic)	_ is part of _ (Part-of)	73
_ is a kind of _ (Taxonomic)	_ is made of _ (Made-of)	37
_ is a kind of _ (Taxonomic)	_ typically performs action _ on _ (Actions)	30
_ is a kind of _ (Taxonomic)	_ is a property of _ (Properties)	25
Grounding	**Central**	**Occurrence**
_ is a kind of _ (Taxonomic)	_ typically performs action _ on _ (Actions)	209
_ is a kind of _ (Taxonomic)	if _ then _ (Conditionals)	202
_ is a kind of _ (Taxonomic)	_ causes _ (Causal)	179
_ is a kind of _ (Taxonomic)	_ changes from _ to _ by _ (Processses)	153
_ is a kind of _ (Taxonomic)	_ uses _ for _ (Functional)	133

Table 2.7: Most reused categories of grounding-grounding and grounding-central inference pairs in WorldTree.

table shows that these patterns emerge through the use of taxonomic knowledge. This suggests that abstraction, intended as the process of going from concrete concepts in the explanandum to high-level concepts in the explanans, is a fundamental part of the inference required for explanation and it is what allows submsuming the explanandum under unifying regularities. Central sentences, in fact, tend to be represented by a more diverse set of linguistic categories in line with those described in the philosophical accounts (i.e., causes, processes, general rules). By looking at grounding-grounding connections, it is possible to notice the relatively high frequency of pairs of taxonomic relations ("_is a kind of_" statements), which confirms again the parallel between explanatory patterns in the corpus and the process of instantiating abstract schematic sentences for unification. Moreover, the presence of linguistic elements about generic attributes and properties is in line with the analysis on the Biology Why Corpus, supporting the fact that these pragmatic elements in natural language explanations play an important role in the abstraction-instantiation process.

Table 2.8 shows examples of sentence-level explanatory patterns, demonstrating how the process of abstraction and unification concretely manifests in the corpus. Table 2.8, in fact, shows that the majority of the grounding-grounding pairs represent inference chains whose function is to perform abstraction (e.g., "an animal is a kind of living thing", "a living thing is a kind of object"). However, we observe that grounding-grounding pairs do not exclusively follow this pattern (e.g., "An animal is a kind of organism", "A plant is a kind of organism"). Specifically, this case suggests that taxonomic relations might play an additional role in the unification process, that is the

Grounding	Grounding	Occurrence
An animal is a kind of living thing	A living thing is a kind of object	18
An animal is a kind of organism	A plant is a kind of organism	14
A human is a kind of animal	An animal is a kind of organism	14
A tree is a kind of plant	A plant is a kind of organism	11
A human is a kind of animal	An animal is a kind of living thing	11

Grounding	Central	Occurrence
Water is a kind of liquid at room temperature	Boiling;evaporation means matter; a substance changes from a liquid into a gas by increasing heat energy	20
Metal is a kind of material	If an object is made of a material then that object has the properties of that material	14
Earth is a kind of planet	A planet rotating causes cycles of day and night on that planet	9
A plant is a kind of organism	Decomposition is when a decomposer breaks down dead organisms	9
Water is a kind of liquid at room temperature	Freezing means matter; a substance changes from a liquid into a solid by decreasing heat energy	9
Metal is a kind of material	Metal is a thermal; thermal energy conductor	9

Table 2.8: Most reused sentence-level grounding-grounding and grounding-central inference pairs in WorldTree.

one of connecting distinct concrete concepts (i.e., "animal" and "plant") to common high-level categories (i.e., "organism").

Overall, it is possible to conclude that explanatory patterns emerging in natural language explanations are closely related to unification, and that this process is fundamentally supported by an inference substrate performing abstraction, whose function is to connect the explanandum to the description of high-level patterns and unifying regularities.

2.3.3 Summary

The main results and findings of the corpus analysis can be summarised as follows:

1. **Natural language explanations are not limited to causes and mechanisms.**

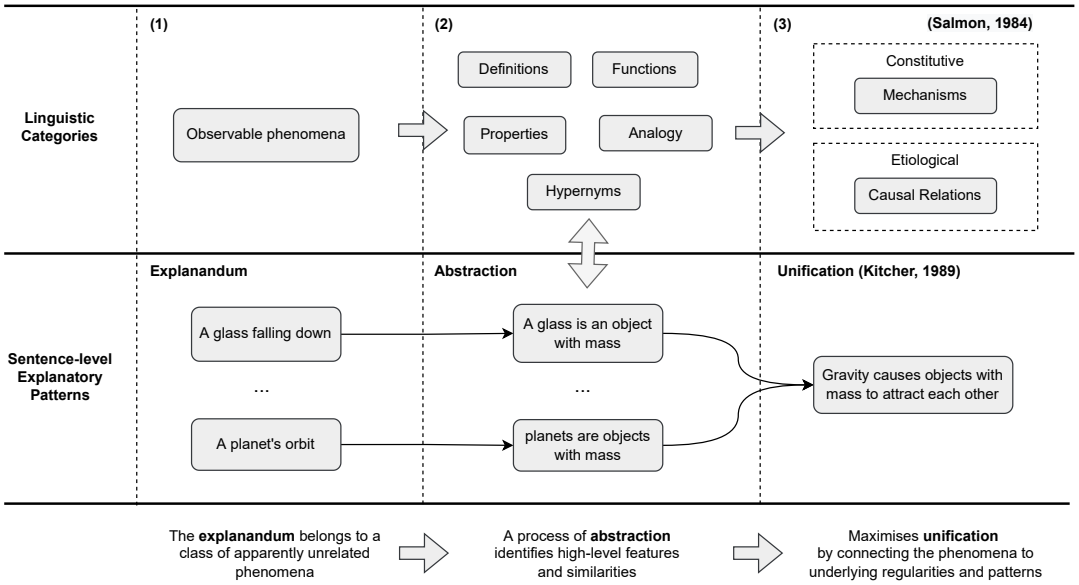


Figure 2.8: A synthesis between the formal accounts of scientific explanations and linguistic aspects found through the corpus analysis.

While *constitutive* and *etiological* elements represent the core part of an explanation, our analysis suggests that additional knowledge categories such as *definitions*, *properties* and *taxonomic relations* play an equally important role in natural language. This can be attributed to both *pragmatic aspects* of explanations and inference requirements associated to *unification*.

2. **Patterns of unification naturally emerge in corpora of explanations.** Even if not intentionally modelled, *unification* seems to be an emergent property of corpora of natural language explanations. The corpus analysis, in fact, reveals that the frequency of reuse of certain explanatory sentences is connected with the notion of *unification power*. Moreover, a qualitative analysis suggests that reused explanatory facts might act as *schematic sentences*, with abstract entities representing the linguistic counterpart of *variables* and *filling instructions* in the Unificationist account.

3. **Unification is realised through a process of abstraction.** Specifically, abstraction represents the fundamental inference substrate supporting unification in natural language. The corpus analysis, in fact, suggests that recurring explanatory patterns emerge through inference chains connecting concrete instances in the explanandum to high-level concepts in the central explanans. This process, realised through specific linguistic elements such as *definitions* and *taxonomic*

relations, is a fundamental part of natural language explanations, and represents what allows subsuming the event to be explained under high-level patterns and unifying regularities.

2.4 Synthesis

Finally, with the help of Figure 2.8, it is possible to perform a synthesis between the epistemological accounts of scientific explanation and the linguistic aspects emerging from the corpus analysis.

In general, explanations cannot be exclusively characterised in terms of *inductive* or *deductive* arguments. This is because of the logical structure of explanations and predictions is intrinsically different [178]. From an epistemic perspective, in fact, the main function of an explanatory argument is to fit the explanandum into a broader pattern that maximises unification, showing that a set of apparently unrelated phenomena are part of a common regularity [90, 91]. From a linguistic point of view, the process of unification tends to generate sentence-level *explanatory patterns* that can be reused and instantiated for deriving and explaining many phenomena. In natural language, unification generally emerges as a process of *abstraction* from the explanandum through the implicit search of common high-level features and similarities between different phenomena.

From an ontic perspective, causal interactions and mechanisms constitute the central part of an explanation as they make the difference between the occurrence and non occurrence of the explanandum [132, 105]. Moreover, causal interactions are responsible for high-level regularities and invariants, with many phenomena being the result of the same underlying causal mechanisms. Here, abstraction represents the inference substrate linking the explanandum to these regularities, a process that manifests in natural language through the use of specific linguistic elements coupled with causes and mechanisms, such as definitions, taxonomic relations, and analogies.

2.5 Implications for Explanation-based NLI

Current lines of research in Explanation-based NLI focus on the development and evaluation of explanation-based models, capable of performing inference through the generation of natural language explanations [173, 179, 72, 154].

Evaluating quality and properties of natural language explanations is still extremely

challenging [69], with most of the existing work focusing on inferential properties in terms of *entailment* or *supporting facts* [187, 15, 31]. This study, however, shows that natural language explanations cannot be evaluated exclusively in terms of deductive reasoning and entailment. This is because deductive arguments cannot fully characterise explanations, and cannot distinguish explanatory arguments from mere predictive ones. As the main function of an explanation is to perform unification, the evaluation methodologies should explicitly reflect this property.

Regarding the construction of explanation-centred corpora, while unification seems to be an emergent property of existing datasets [179, 72], future research can benefit from explicitly considering it during the annotation process. Unification patterns, in fact, can provide a top-down and reuse-oriented methodology to facilitate evaluation and scale up the annotation process. From an inferential perspective, the evaluation of natural language explanations should focus on a multi-dimensional set of inference capabilities, assessing explanation-based systems in the ability to perform abstraction, identify unifying causal mechanisms and interpret high-level regularities. Finally, emergent unification patterns in natural language explanations can provide a way to build more robust inference models.

2.6 Conclusion

In order to provide an epistemologically grounded characterisation of natural language explanations, this chapter attempted to bridge the gap in the notion of *scientific explanation* [131, 132], studying it as both a *formal object* and as a *linguistic expression*. The combination of a systematic survey with a corpus analysis on natural language explanations [70, 72], allowed us to derive specific conclusions on the nature of explanatory arguments from both a top-down (categorical) and a bottom-up (corpus-based) perspective:

1. Explanations cannot be entirely characterised in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*.

2. A scientific explanation must cite causes and mechanisms that are responsible for the occurrence of the explanandum.

3. While natural language explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms.

4. Patterns of unification naturally emerge in corpora of explanations even if not intentionally modelled.

5. Unification emerges through a process of abstraction, whose function is to provide the inference support for subsuming the event to be explained under recurring patterns and regularities.

From these findings, it is possible to derive a set of guidelines for future research on Explanation-based NLI for the creation and evaluation of models that can interpret and generate natural language explanations:

1. Explainability cannot be evaluated only in terms of deductive inference capabilities and entailment properties. This is because deductive arguments cannot entirely characterise explanations, and cannot be used to distinguish explanatory arguments from mere predictive ones.

2. As the main function of an explanatory arguments is to perform unification, the evaluation of explainability must explicitly take into account this property. Moreover, while unification seems to be an emergent property of existing benchmarks, it should be explicitly considered as a top-down approach for the creation of explanation-centred corpora to facilitate evaluation.

3. From a bottom-up perspective, the evaluation of explainability should not only focus on specific inference properties connected to causality, but also take into account other features of explanation, including semantic abstraction and analogy making.

4. The unification property of explanatory arguments can provide a way to build more robust inference models that explicitly leverage patterns of derivation, as well as more efficient and scalable solutions to construct explanation-centred corpora. Recurring argument patterns, in fact, can potentially reduce the search space for multi-hop inference models and support a more schematic, reuse-oriented mechanism for the annotation of gold explanations.

The chapter contributed to addressing a fundamental gap in classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts. This characterisation can support a more principled design and evaluation of Explanation-based NLI systems which can better interpret and generate natural language explanations.

2.7 Scoping and Limitations

The survey focused on epistemological accounts that attempt to define an *objective* relationship between *explanandum* and *explanans*. While characterising explanatory arguments is important for a complete understanding of the concept, explanation is a broader topic that embraces different aspects not considered in the survey, such as cognitive processes, conversational acts, as well as pragmatic and contextual elements involved in humans' communication [113]. While these aspects might be relevant for the construction of Explanation-based NLI models, they were considered out-of-scope for the thesis and left as a possible focus for future work. Regarding the surveyed accounts, while some consensus on the nature and function of explanation exists, philosophers still disagree on whether the discussed features apply to all types of scientific explanations and are transferable across different fields and domains [131]. Therefore, additional work is still required to derive a complete and universally accepted account and investigate whether the considered features are suitable for a general description of explanations.

While relating the corpus analysis to epistemological accounts allows drawing conclusions that are generalisable to some extent, the presented quantitative methodology relies on specific features of the analysed resources. Specifically, the discussed method adopted to investigate the emergence of patterns of unification could only be applied on corpora with a reuse-oriented design such as WorldTree. With the current methodology, in fact, it is not yet clear how to possibly identify such patterns through the re-occurrence of specific facts in corpora that do not possess this property.

Chapter 3

Unification-based Inference

This chapter aims to investigate ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"* proposing a novel Explanation Regeneration framework for science questions. Following the study presented in Chapter 2, this chapter describes a method to leverage *explicit explanatory patterns* in corpora of scientific explanations emerging in the form of *unification*. Specifically, the unification-based framework ranks atomic sentences in an external fact bank, estimating their *explanatory relevance* via the integration of lexical relevance metrics and the notion of *explanatory power*, computed analysing explanations for similar questions in the corpus.

An extensive evaluation is performed integrating k-NN clustering and sparse Information Retrieval (IR) techniques. The chapter presents the following conclusions: (1) The proposed method achieves results competitive with some of the existing Transformer-based models, yet being orders of magnitude faster (2) The unification-based mechanism has a key role in reducing semantic drift, contributing to the reconstruction of long explanations (6 or more facts) and the ranking of complex inference facts (+12.0 Mean Average Precision) (3) The constructed explanations can support downstream QA models, improving the accuracy of a BERT baseline by up to 10% overall[1].

[1]This chapter follows the publication "Unification-based Reconstruction of Multi-hop Explanations for Science Questions"[161]

3.1 Introduction

Answering *multiple-choice science questions* has become an established benchmark for testing natural language understanding and complex inference [86, 22, 111]. In parallel with other research areas in NLI, a crucial requirement emerging in recent years is *explainability* [151, 113, 12, 127]. To improve development and evaluation of automatic methods of inference, in fact, it is necessary not only to measure performance on downstream answer prediction, but also the ability of a NLI system to provide explanations for the underlying reasoning process.

The need for explainability and a quantitative methodology for its evaluation have conducted to the creation of shared tasks on *Explanation Regeneration* [71] using corpora of explanations such as Worldtree [72, 67]. Given a science question, Explanation Regeneration consists in reconstructing the gold explanation that supports the correct answer through multi-hop inference on a series of atomic facts. While most of the existing benchmarks for multi-hop inference require the composition of only 2 supporting sentences or paragraphs (e.g. QASC [86], HotpotQA [187]), Explanation Regeneration on science questions requires the aggregation of an average of 6 facts (and as many as $\approx$20), making it particularly hard for existing models. Moreover, the structure of the explanations affects the complexity of the regeneration task. Explanations for science questions are typically composed of two main parts: a grounding part, containing knowledge about concrete concepts in the question, and a core scientific part, describing general laws and regularities.

Consider the following question and answer pair from WorldTree [72]:

- q: what is an example of a **force** producing heat?
  a: two **sticks** getting warm when **rubbed together**.

An explanation that justifies a is composed using the following sentences from the corpus: *(f_1) a **stick** is a kind of object*; *(f_2) to **rub together** means to move against*; *(f_3) friction is a kind of **force***; *(f_4) friction occurs when two objects' surfaces move against each other*; *(f_5) friction causes the temperature of an object to increase*. The explanation contains a set of concrete sentences that are conceptually connected with q and a (f_1,f_2 and f_3), along with a set of abstract facts that require multi-hop inference (f_4 and f_5).

Previous work has shown that constructing long explanations is challenging due to *semantic drift* – i.e. the tendency of composing out-of-context inference chains as the number of hops increases [82, 46]. While existing approaches build explanations

considering each question in isolation [85, 89], we hypothesise that semantic drift can be tackled by leveraging *explanatory patterns* emerging in clusters of similar questions.

In science, a given statement is considered explanatory to the extent it performs *unification* [47, 90, 91], that is, showing how a set of initially disconnected phenomena are the expression of the same regularity. Since the explanatory power of a given statement depends on the number of unified phenomena, highly explanatory facts tend to create *unification patterns* – i.e. similar phenomena require similar explanations. Coming back to our example, we hypothesise that the relevance of abstract statements requiring multi-hop inference, such as f_4 (*"friction occurs when two objects' surfaces move against each other"*), can be estimated considering its unification power.

Following these observations, we present a framework that ranks atomic facts through the combination of two scoring functions:

- A *Relevance Score (RS)* that represents the lexical relevance of a given fact.
- An *Explanatory Power (PW)* score that models the unification power of a fact according to its frequency in explanations for similar questions.

An extensive evaluation is performed on the WorldTree corpus [72, 71], adopting a combination of k-NN clustering and Information Retrieval (IR) techniques. We present the following conclusions:

1. Despite its simplicity, the proposed method achieves results competitive with some of the existing Transformers-based models [34, 21], yet being orders of magnitude faster, a feature that makes it scalable to large explanation-based corpora.
2. We empirically demonstrate the key role of the unification-based mechanism in the regeneration of long explanations (6 or more facts) and explanations requiring complex inference (+12.0 Mean Average Precision).
3. Crucially, the constructed explanations can support downstream question answering models, improving the accuracy of a BERT baseline [39] by up to 10% overall.

To the best of our knowledge, we are the first to propose a method that leverages unification patterns for the regeneration of multi-hop explanations, and empirically demonstrate their impact on semantic drift and downstream question answering.

3.2 Explanation Regeneration as a Ranking Problem

A multiple-choice science question $Q = \{q, C\}$ is a tuple composed by a question q and a set of candidate answers $C = \{c_1, c_2, \ldots, c_n\}$. Given an hypothesis h_j defined as the concatenation of q with a candidate answer $c_j \in C$, the task of Explanation Regeneration consists in selecting a set of atomic facts from a knowledge base $E_j = \{f_1, f_2, \ldots, f_n\}$ that support and justify h_j.

In this paper, we adopt a methodology that relies on the existence of a corpus of explanations. A corpus of explanations is composed of two distinct knowledge sources:

- A primary knowledge base, *Facts KB* (F_{kb}), defined as a collection of sentences $F_{kb} = \{f_1, f_2, \ldots, f_n\}$ encoding the general world knowledge necessary to answer and explain science questions. A fundamental and desirable characteristic of F_{kb} is *reusability* – i.e. each of its facts f_i can be potentially reused to compose explanations for multiple questions

- A secondary knowledge base, *Explanation KB* (E_{kb}), consisting of a set of tuples $E_{kb} = \{(h_1, E_1), (h_2, E_2), \ldots, (h_m, E_m)\}$, each of them connecting a true hypothesis h_j to its corresponding explanation $E_j = \{f_1, f_2, \ldots, f_k\} \subseteq F_{kb}$. An explanation $E_j \in E_{kb}$ is therefore a composition of facts belonging to F_{kb}.

In this setting, the Explanation Regeneration task for an unseen hypothesis h_j can be modelled as a ranking problem [71]. Specifically, given an hypothesis h_j the algorithm to solve the task is divided into three macro steps:

1. Computing an explanatory score $s_i = e(h_j, f_i)$ for each fact $f_i \in F_{kb}$ with respect to h_j

2. Producing an ordered set $Rank(h_j) = \{f_1, \ldots, f_k, f_{k+1}, \ldots, f_n \mid s_k \geq s_{k+1}\} \subseteq F_{kb}$

3. Selecting the top k elements belonging to $Rank(h_j)$ and interpreting them as an explanation for h_j; $E_j = topK(Rank(h_j))$.

3.3 Modelling Explanatory Relevance

To investigate ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"* for Explanation Regeneration, we present an approach for modelling $e(h_j, f_i)$ that is guided by the following research hypotheses:

- **RH3.1**: Scientific explanations are composed of a set of concrete facts connected to the question, and a set of abstract statements expressing general scientific laws and regularities. Concrete facts tend to share key concepts with the question and can therefore be effectively ranked by IR techniques based on lexical relevance.

- **RH3.2**: General scientific statements tend to be abstract and therefore difficult to rank by means of lexical relevance. However, due to explanatory unification, core scientific statements tend to create explicit patterns by being frequently reused across similar questions. We hypothesise that these patterns can be captured through a model of explanatory power, whose value is proportional to the number of times a given fact f_i explains clusters of similar questions.

To formalise these research hypotheses, we model the explanatory scoring function $e(h_j, f_i)$ as a combination of two components:

$$e(h_j, f_i) = \lambda \cdot rs(h_j, f_i) + (1-\lambda) \cdot pw(h_j, f_i) \tag{3.1}$$

Here, $rs(h_j, f_i)$ represents a lexical Relevance Score (RS) assigned to $f_i \in F_{kb}$ with respect to h_j, while $pw(h_j, f_i)$ represents the Explanatory Power (PW) of f_i computed over E_{kb} as follows:

$$pw(h_j, f_i) = \sum_{h_z \in kNN(h_j)}^{K} sim(h_j, h_z) \cdot \mathbb{1}(f_i, h_z) \tag{3.2}$$

$$\mathbb{1}(f_i, h_z) = \begin{cases} 1 & \text{if } f_i \in E_z \\ 0 & \text{if } f_i \notin E_z \end{cases} \tag{3.3}$$

$kNN(h_j) = \{(h_1, E_1), \ldots (h_k, E_k)\} \subseteq E_{kb}$ is the set of k-nearest neighbours of h_j belonging to E_{kb} retrieved according to a similarity function $sim(h_j, h_z)$. On the other hand, $\mathbb{1}(f_i, h_z)$ is the indicator function verifying whether the fact f_i is included in the explanation E_z for the hypothesis h_z.

In the formulation of Equation 3.2 we aim to capture two main aspects related to our research hypotheses:

1. The more a fact f_i is reused for explanations in E_{kb}, the higher its explanatory power;

2. The explanatory power of a fact f_i is proportional to the similarity between the hypotheses in E_{kb} that are explained by f_i and the unseen hypothesis (h_j) we want to explain.

Figure 3.1 shows a schematic representation of the Unification-based framework.

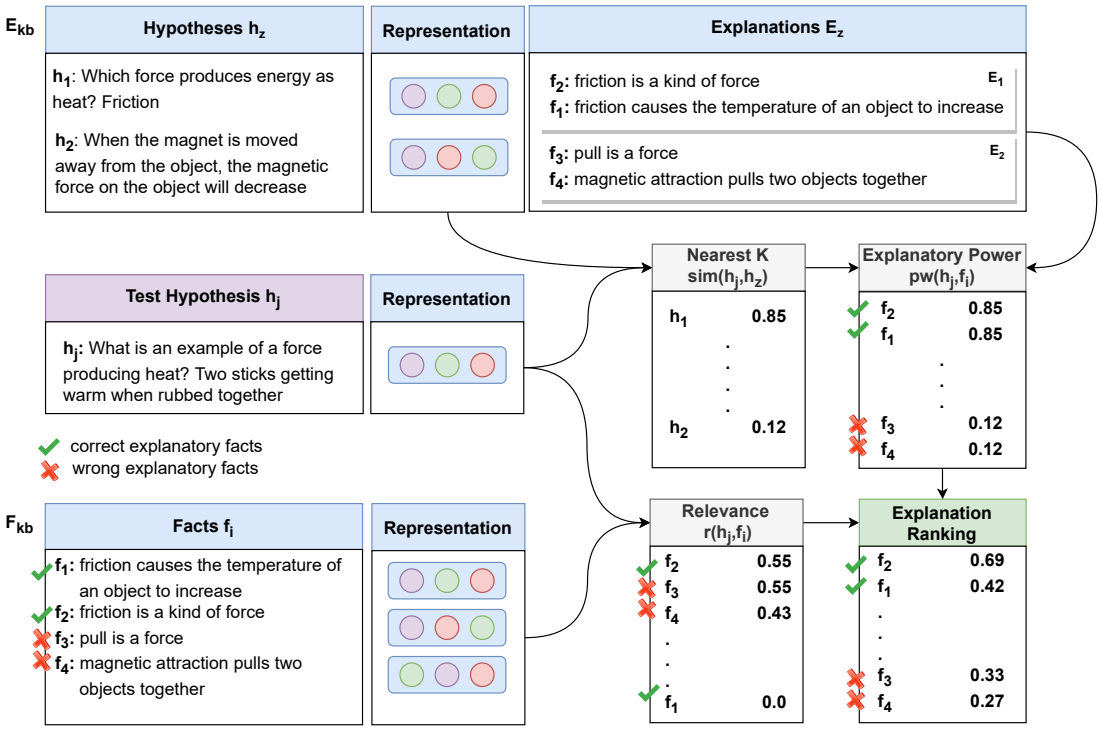


Figure 3.1: Overview of the Unification-based framework for Explanation Regeneration.

3.4 Empirical Evaluation

We carried out an empirical evaluation on the WorldTree corpus [72], a subset of the ARC dataset [22] that includes explanations for science questions. The corpus provides an explanation-based knowledge base (F_{kb} and E_{kb}) where each explanation in E_{kb} is represented as a set of lexically connected sentences describing how to arrive at the correct answer. The science questions in the WorldTree corpus are split into *training-set*, *dev-set*, and *test-set*. The gold explanations in the *training-set* are used to build the Explanation KB (E_{kb}), while the gold explanations in *dev* and *test* set are used for *evaluation purpose only*. The corpus groups the explanation sentences belonging to E_{kb} into three explanatory roles: *grounding*, *central* and *lexical glue*.

Consider the example in Figure 3.1. To support q and c_j the system has to retrieve the scientific facts describing how friction occurs and produces heat across objects. The corpus classifies these facts (f_1) as *central*. *Grounding* explanations like *"friction is a kind of force"* (f_2) link question and answer to the central explanations. *Lexical glues* such as *"to rub; to rub together means to mover against"* are used to fill lexical gaps between sentences. Additionally, the corpus divides the facts belonging to F_{kb} into three inference categories: *retrieval type*, *inference supporting type*, and *complex*

inference type. Taxonomic knowledge and properties such as *"friction is a kind of force"* (f_2) are classified as *retrieval type*. Facts describing actions, affordances, and requirements such as *"friction occurs when two object's surfaces move against each other"* are grouped under the *inference supporting types*. Knowledge about causality, description of processes and if-then conditions such as *"friction causes the temperature of an object to increase"* (f_1) is classified as *complex inference*.

We implement Relevance and Explanatory Power adopting TF-IDF/BM25 vectors and cosine similarity function (i.e. $1 - cos(\vec{x}, \vec{y})$). Specifically, The RS model uses TF-IDF/BM25 to compute the relevance function for each fact in F_{kb} (i.e. $rs(h_j, f_i)$ function in Equation 3.1) while the PW model adopts TF-IDF/BM25 to assign similarity scores to the hypotheses in E_{kb} (i.e. $sim(h_j, h_z)$ function in Equation 3.2). For reproducibility, the code is available online[2]. Additional details can be found in Appendix A.

3.4.1 Explanation Regeneration

In line with the shared task [71], the performances of the models are evaluated via Mean Average Precision (MAP) of the explanation ranking produced for a given question q_j and its correct answer a_j.

Table 3.1 illustrates the score achieved by our best implementation compared to state-of-the-art approaches in the literature. Previous approaches are grouped into four categories: *Transformers*, *Information Retrieval with re-ranking*, *One-step Information Retrieval*, and *Feature-based models*.

Transformers. This class of approaches employs the gold explanations in the corpus to train a BERT language model [39]. The best-performing system [34] adopts a multi-step retrieval strategy. In the first step, it returns the top K sentences ranked by a TF-IDF model. In the second step, BERT is used to re-rank the paths composed of all the facts that are within 1-hop from the first retrieved set. Similarly, other approaches adopt BERT to re-rank each fact individually [6, 21].

Although the best model achieves state-of-the-art results in Explanation Regeneration, these approaches are computationally expensive, being limited by the application of a pre-filtering step to contain the space of candidate facts. Consequently, these systems do not scale with the size of the corpus. We estimated that the best performing model [34] takes $\approx$ 10 hours to run on the whole test set (1240 questions) using 1 Tesla 16GB V100 GPU.

[2] `https://github.com/ai-systems/unification_regeneration_explanations`

Model	Description	Trained	MAP	
			Test	Dev
Transformers				
Das et al. [34]	BERT re-ranking with inference chains	Yes	**56.3**	**58.5**
Chia et al. [21]	BERT re-ranking with gold IR scores	Yes	47.7	50.9
Banerjee [6]	BERT iterative re-ranking	Yes	41.3	42.3
IR with re-ranking				
Chia et al. [21]	Iterative BM25	**No**	45.8	49.7
One-step IR				
BM25	BM25 Relevance Score	**No**	43.0	46.1
TF-IDF	TF-IDF Relevance Score	**No**	39.4	42.8
Feature-based				
D'Souza et al.[30]	Feature-rich SVM ranking + Rules	Yes	39.4	44.4
D'Souza et al. [30]	Feature-rich SVM ranking	Yes	34.1	37.1
Unification-based				
RS + Pw (Best)	Joint Relevance and Explanatory Power	**No**	**50.8**	**54.5**
Pw (Best)	Explanatory Power	**No**	22.9	21.9

Table 3.1: Results on test and dev set and comparison with state-of-the-art approaches. The column **trained** indicates whether the model requires an explicit training session on the Explanation Regeneration task.

Comparatively, our model constructs explanations for all the questions in the test set in $\approx$ 30 seconds, without requiring the use of GPUs ($<$ 1 second per question). This feature makes the Unification-based regeneration suitable for large corpora and downstream question answering models (as shown in Section 3.4.4). Moreover, our approach does not require any explicit training session on the Explanation Regeneration task, with significantly reduced number of parameters to tune. Along with scalability, the proposed approach achieves results comparable with some of the existing Transformers-based models (50.8/54.5 MAP). Although we observe lower performance when compared to the best-performing approach (-5.5/-4.0 MAP), the joint RS + PW model outperforms two BERT-based models [21, 6] on both test and dev set by 3.1/3.6 and 9.5/12.2 MAP, respectively.

Information Retrieval with re-ranking. Chia et al. [21] describe a multi-step, iterative re-ranking model based on BM25. The first step consists in retrieving the explanation sentence that is most similar to the question adopting BM25 vectors. During the second step, the BM25 vector of the question is updated by aggregating it with the retrieved explanation sentence vector through a `max` operation. The first and second steps are repeated for K times. Although this approach uses scalable IR techniques, it relies on a multi-step retrieval strategy. Besides, the RS + PW model outperforms this approach on both test and dev set by 5.0/4.8 MAP, respectively.

One-step Information Retrieval. We compare the RS + PW model with two IR baselines. The baselines adopt TF-IDF and BM25 to compute the Relevance Score only – i.e. the $us(q, c_j, f_i)$ term in Equation 1 is set to 0 for each fact $f_i \in F_{kb}$. In line with previous IR literature [129], BM25 leads to better performance than TF-IDF. While these approaches share similar characteristics, the combined RS + PW model outperforms both RS BM25 and RS TF-IDF on test and dev-set by 7.8/8.4 and 11.4/11.7 MAP. Moreover, the joint RS + PW model improves the performance of the PW model alone by 27.9/32.6 MAP. These results outline the complementary aspects of Relevance and Explanatory Power. We provide a detailed analysis by performing an ablation study on the dev-set (Section 3.4.2).

Feature-based models. D'Souza et al. [30] propose an approach based on a learning-to-rank paradigm. The model extracts a set of features based on overlaps and coherence metrics between questions and explanation sentences. These features are then given as

Model	MAP			
	All	Central	Grounding	Lexical Glue
RS TF-IDF	42.8	43.4	25.4	8.2
RS BM25	46.1	46.6	23.3	10.7
PW TF-IDF	21.6	16.9	22.0	13.4
PW BM25	21.9	18.1	16.7	15.0
RS TF-IDF + PW TF-IDF	48.5	46.4	**32.7**	11.7
RS TF-IDF + PW BM25	50.7	48.6	30.42	13.4
RS BM25 + PW TF-IDF	51.9	48.2	31.7	14.8
RS BM25 + PW BM25	**54.5**	**51.7**	27.3	**16.7**

(a) Explanatory roles.

Model	MAP		
	1+ Overlaps	1 Overlap	0 Overlaps
RS TF-IDF	57.2	33.6	7.1
RS BM25	62.2	37.1	7.1
PW TF-IDF	17.37	18.0	12.5
PW BM25	18.1	18.1	**13.1**
RS TF-IDF + PW TF-IDF	60.2	38.4	9.0
RS TF-IDF + PW BM25	62.5	39.5	9.6
RS BM25 + PW TF-IDF	61.3	40.6	11.0
RS BM25 + PW BM25	**64.8**	**41.9**	11.2

(b) Lexical overlaps with the hypothesis.

Model	MAP		
	Retrieval	Inference-supporting	Complex Inference
RS TF-IDF	33.5	34.7	21.8
RS BM25	36.0	36.1	24.8
PW TF-IDF	17.6	12.8	19.5
PW BM25	16.8	13.2	20.9
RS TF-IDF + PW TF-IDF	38.3	33.2	30.2
RS TF-IDF + PW BM25	40.0	35.6	33.3
RS BM25 + PW TF-IDF	40.5	33.6	33.4
RS BM25 + PW BM25	**40.6**	**38.3**	**36.8**

(c) Inference types.

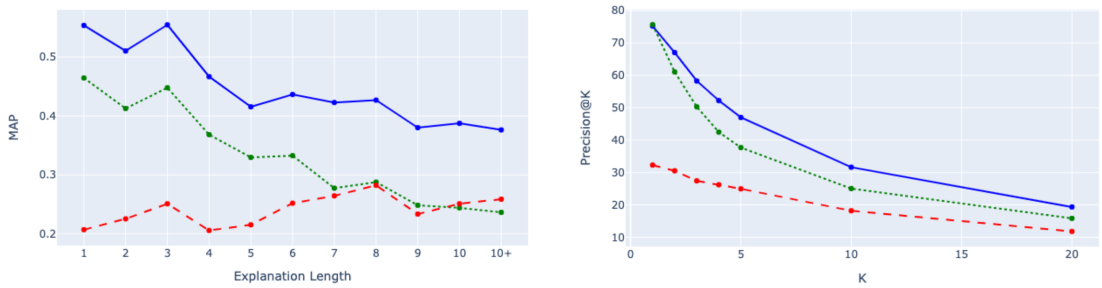
Table 3.2: Detailed analysis of the performance (dev-set) by breaking down the gold explanatory facts according to their explanatory role (2.a), number of lexical overlaps with the question (2.b) and inference type (2.c).

input to a SVM ranker module. While this approach scales to the whole corpus without requiring any pre-filtering step, it is significantly outperformed by the RS + PW model on both test and dev set by 16.7/17.4 MAP, respectively.

3.4.2 Explanation Analysis

We present an ablation study with the aim of understanding the contribution of each sub-component to the general performance of the joint RS + PW model (see Table 3.1). To this end, a detailed evaluation on the development set of the WorldTree corpus is carried out, analysing the performance in reconstructing explanations of different types and complexity. We compare the joint model (RS + PW) with each individual sub-component (RS and PW alone). In addition, a set of qualitative examples are analysed to provide additional insights on the complementary aspects captured by Relevance and Explanatory Power.

Explanatory categories. Given a question q_j and its correct answer a_j, we classify a fact f_i belonging to the gold explanation E_j according to its explanatory role (*central, grounding, lexical glue*) and inference type (*retrieval, inference-supporting and complex*



(a) MAP vs Explanation length.

(b) Precision@K.

Figure 3.2: Impact of the Explanatory Power on semantic drift (3.a) and precision (3.b). RS + PW (Blue Straight), RS (Green Dotted), PW (Red Dashed).

inference). In addition, three new categories are derived from the number of overlaps between f_i and the concatenation of q_j with a_j (h_j) computed by considering nouns, verbs, adjectives and adverbs (1+ overlaps, 1 overlap, 0 overlaps).

Table 2 reports the MAP score for each of the described categories. Overall, the best results are obtained by the BM25 implementation of the joint model (RS BM25 + PW BM25) with a MAP score of 54.5. Specifically, RS BM25 + PW BM25 achieves a significant improvement over both RS BM25 (+8.5 MAP) and PW BM25 (+32.6 MAP) baselines. Regarding the explanatory roles (Table 3.2a), the joint TF-IDF implementation shows the best performance in the regeneration of *grounding* explanations (32.7 MAP). On the other hand, a significant improvement over the RS baseline is obtained by RS BM25 + PW BM25 on both *lexical* glues and *central* explanation sentences (+6.0 and +5.6 MAP over RS BM25).

Regarding the lexical overlaps categories (Table 3.2b), we observe a steady improvement for all the combined RS + PW models over the respective RS baselines. Notably, the PW models achieve the best performance on the 0 overlaps category, which includes the most challenging facts for the RS models. The improved ability to rank abstract explanatory facts contributes to better performance for the joint models (RS + PW) in the regeneration of explanations that share few terms with question and answer (*1 Overlap* and *0 Overlaps* categories). This characteristic leads to an improvement of 4.8 and 4.1 MAP for the RS BM25 + PW BM25 model over the RS BM25 baseline.

Similar results are achieved on the inference types categories (Table 3.2c). Crucially, the largest improvement is observed for *complex inference* sentences where RS BM25 + PW BM25 outperforms RS BM25 by 12.0 MAP, confirming the decisive contribution of the Explanatory Power to the ranking of complex scientific facts.

Semantic drift. Science questions in the WorldTree corpus require an average of six facts in their explanations [67]. Long explanations typically include sentences that share few terms with question and answer, increasing the probability of semantic drift. Therefore, to test the impact of the Explanatory Power on the robustness of the model, we measure the performance in the regeneration of many-hops explanations.

Figure 5.3b shows the change in MAP score for the RS + PW, RS and PW models (BM25) with increasing explanation length. The fast drop in performance for the Relevance Score reflects the complexity of the task. This drop occurs because the RS model is not able to rank abstract explanatory facts. Conversely, the PW model exhibits increasing performance, with a trend that is inverse. Short explanations, indeed, tend to include question-specific facts with low explanatory power. On the other hand, the longer the explanation, the higher the number of core scientific facts. Therefore, the decrease in MAP observed for the RS model is compensated by the Explanatory Power, since core scientific facts tend to form unification patterns across similar questions. This results demonstrate that the Explanatory Power has a crucial role in alleviating the semantic drift for the joint model (RS + PW), resulting in a larger improvement on many-hops explanations (6+ facts).

Similarly, Figure 3.2b illustrates the Precision@K. As shown in the graph, the drop in precision for the PW model exhibits the slowest degradation. Similarly to what observed for many-hops explanations, the PW score contributes to the robustness of the RS + PW model, making it able to reconstruct more precise explanations. As discussed in section 3.4.4, this feature has a positive impact on question answering.

k-NN clustering. We investigate the impact of the k-NN clustering on the Explanation Regeneration task. Figure 3.3 shows the MAP score obtained by the joint RS + PW model (BM25) with different numbers k of nearest hypotheses considered for the Explanatory Power. The graph highlights the improvement in MAP achieved with increasing values of k. Specifically, we observe that the best MAP is obtained with $k = 100$. These results confirm that the explanatory power can be effectively estimated using clusters of similar hypotheses, and that the unification-based mechanism has a crucial role in improving the performance of the relevance model.

3.4.3 Qualitative analysis

To provide additional insights on the complementary aspects of Explanatory Power and Relevance Score, we present a set of qualitative examples from the dev-set. Table 3.3

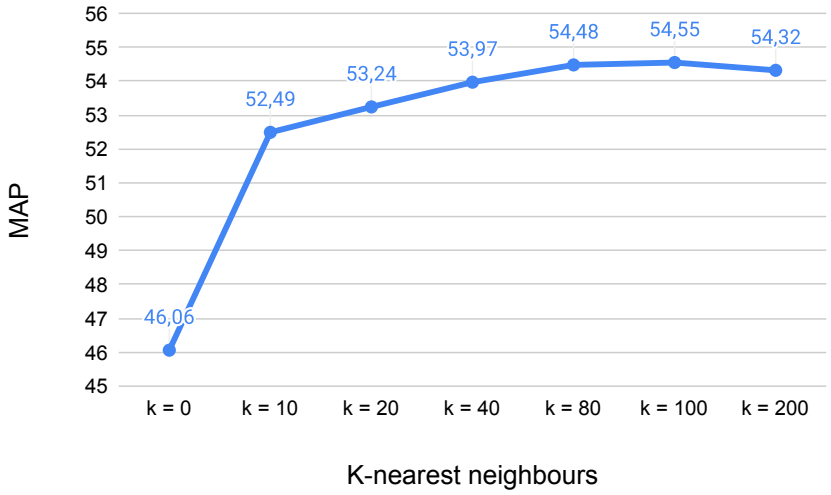


Figure 3.3: Impact of the k-NN clustering on the final MAP score. The value k represents the number of similar hypotheses considered for the Explanatory Power.

illustrates the ranking assigned by RS and RS + PW models to scientific sentences of increasing complexity. The words in **bold** indicate lexical overlaps between question, answer and explanation sentence. In the first example, the sentence *"gravity; gravitational force causes objects that have mass; substances to be pulled down; to fall on a planet"* shares key terms with question and candidate answer and is therefore relatively easy to rank for the RS model (#36). Nevertheless, the RS + PW model is able to improve the ranking by 34 positions (#2), as the gravitational law represents a scientific pattern with high explanatory unification, frequently reused across similar questions. The impact of the Explanatory Power is more evident when considering abstract explanatory facts. Coming back to our original example (i.e. *"What is an example of a force producing heat?"*), the fact *"friction causes the temperature of an object to increase"* has no significant overlaps with question and answer. Thus, the RS model ranks the gold explanation sentence in a low position (#1472). However, the Explanatory Power (PW) is able to capture the explanatory power of the fact from similar hypotheses in E_{kb}, pushing the RS + PW ranking up to position #21 (+1451).

3.4.4 Question Answering

To understand whether the constructed explanations can support question answering, we compare the performance of BERT for multiple-choice QA [39] without explanations with the performance of BERT provided with the top K explanation sentences retrieved

Question	Answer	Explanation Fact	Most Similar Hypotheses in E_{kb}	RS	RS + PW
If you bounce a rubber ball on the floor it goes up and then comes down. What **causes** the ball to come down?	**gravity**	**gravity**; gravitational force **causes** objects that have mass; substances to be pulled down; to fall on a planet	(1) A ball is tossed up in the air and it comes back down. The ball comes back down because of - gravity (2) A student drops a ball. Which force causes the ball to fall to the ground? - gravity	#36	#2 (↑34)
Which **animals** would most likely be helped by flood in a coastal area?	alligators	as water increases in an environment, the population of aquatic **animals** will increase	(1) Where would animals and plants be most affected by a flood? - low areas (2) Which change would most likely increase the number of salamanders? - flood	#198	#57 (↑141)
What is an example of a force producing heat?	two sticks getting warm when rubbed together	friction causes the temperature of an object to increase	(1) Rubbing sandpaper on a piece of wood produces what two types of energy? - sound and heat (2) Which force produces energy as heat? - friction	#1472	#21 (↑1451)

Table 3.3: Impact of the Explanatory Power on the ranking of scientific facts with increasing complexity.

by RS and RS + PW models (BM25). BERT without explanations operates on question and candidate answer only. On the other hand, BERT with explanation receives the following input: the question (q), a candidate answer (c_i) and the explanation for c_i (E_i). In this setting, the model is fine-tuned for binary classification ($bert_b$) to predict a set of probability scores $P = \{p_1,\ p_2,\ ...,\ p_n\}$ for each candidate answer in $C = \{c1,\ c_2,\ ...,\ c_n\}$:

$$bert_b(\texttt{[CLS]} \mid\mid q \mid\mid c_i \mid\mid \texttt{[SEP]} \mid\mid E_i) = p_i \tag{3.4}$$

The binary classifier operates on the final hidden state corresponding to the `[CLS]` token. To answer the question q, the model selects the candidate answer c_a such that $a = \text{argmax}_i\, p_i$.

Table 3.4 reports the accuracy with and without explanations on the WorldTree *test-set* for *easy* and *challenge* questions [22]. Notably, a significant improvement in accuracy can be observed when BERT is provided with explanations retrieved by the regeneration modules (+9.84% accuracy with RS BM25 + PW BM25 model). The improvement is consistent on the *easy* split (+6.92%) and particularly significant for *challenge* questions (+15.69%). Overall, we observe a correlation between more precise explanations and accuracy in answer prediction, with BERT + RS being outperformed by BERT + RS + PW for each value of K. The decrease in accuracy occurring with increasing values of K is coherent with the drop in precision for the models observed in Figure 3.2b. Moreover, steadier results adopting the RS + PW model suggest a positive

Model	Accuracy		
	Easy	Challenge	Overall
BERT (no explanation)	48.54	26.28	41.78
BERT + RS (K = 3)	53.20	40.97	49.39
BERT + RS (K = 5)	54.36	38.14	49.31
BERT + RS (K = 10)	32.71	29.63	31.75
BERT + RS + PW (K = 3)	**55.46**	**41.97**	**51.62**
BERT + RS + PW (K = 5)	54.48	39.43	50.12
BERT + RS + PW (K = 10)	48.66	37.37	45.14

Table 3.4: Performance of BERT on question answering (test-set) with and without the Explanation Regeneration models.

contribution from abstract explanatory facts. Additional investigation of this aspect will be a focus for future work.

3.5 Related Work

Explanations for Science Questions. Reconstructing explanations for science questions can be reduced to a multi-hop inference problem, where multiple pieces of evidence have to be aggregated to arrive at the final answer [151, 85, 89, 68]. Aggregation methods based on lexical overlaps and explicit constraints suffer from *semantic drift* [82, 46] – i.e. the tendency of composing spurious inference chains leading to wrong conclusions. One way to contain semantic drift is to leverage common explanatory patterns in explanation-centred corpora [72]. Transformers [34, 21] represent the state-of-the-art for Explanation Regeneration in this setting [71]. However, these models require high computational resources that prevent their applicability to large corpora. On the other hand, approaches based on IR techniques are readily scalable. The approach described in this paper preserves the scalability of IR methods, obtaining, at the same time, performances competitive with some of the existing Transformers. Thanks to this feature, the framework can be flexibly applied in combination with downstream question answering models. Our findings are in line with previous work in different QA settings [123, 185], which highlights the positive impact of explanations and supporting facts on the final answer prediction task. In parallel with Science QA, the development of models for explanation generation is being explored in different NLP tasks, ranging from open domain question answering [187, 155], to textual entailment [15] and natural

language premise selection [45, 44].

Scientific Explanation and AI. The field of Artificial Intelligence has been historically inspired by models of explanation in Philosophy of Science [149]. The deductive-nomological model proposed by Hempel [56] constitutes the philosophical foundation for explainable models based on logical deduction, such as Expert Systems [95, 172] and Explanation-based Learning [116]. Similarly, the inherent relation between explanation and causality [176, 132] has inspired computational models of causal inference [119]. The view of explanation as unification [47, 90, 91] is closely related to Case-based reasoning [92, 142, 36]. In this context, analogical reasoning plays a key role in the process of reusing abstract patterns for explaining new phenomena [148]. Similarly to our approach, Case-based reasoning applies this insight to construct solutions for novel problems by retrieving, reusing and adapting explanations for known cases solved in the past.

3.6 Conclusion

This paper proposed a novel framework for multi-hop Explanation Regeneration based on *explanatory unification*. An extensive evaluation on the WorldTree corpus led to the following conclusions: (1) The approach is competitive with some of the existing Transformers-based models, yet being significantly faster and inherently scalable; (2) The unification-based mechanism supports the ranking of complex and long explanations; (3) The constructed explanations improves the accuracy of a BERT baseline for question answering by up to 10% overall.

3.7 Scoping and Limitations

The model of explanatory power presented in Chapter 3 relies on the availability of human-annotated explanations with specific features (e.g., explanatory facts reused across different training instances). However, these resources might not be available in real-world scenarios and are generally costly to develop. Moreover, since the explanatory power model relies on similarities measures, the model's ability to generalise might be susceptible to the incompleteness of the facts bank and the availability of representative explanations. Finally, due to the use of the indicator function, the current implementation of the model is not able to identify sentences in the facts bank that have

different surface forms but same underlying meaning, preventing the ability to estimate the explanatory power of sentences that are not explicitly used in the gold explanations.

Chapter 4

Case-based Abductive NLI

This chapter keeps investigating ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"* focusing on downstream Abductive Natural Language Inference. Specifically, the chapter proposes an abductive framework for Explanation-based NLI built upon the *retrieve-reuse-refine* paradigm in *case-based reasoning*. Following the notion of explanatory power discussed in Chapter 3, the case-based reasoning model attempts to explain unseen natural language hypotheses by retrieving and adapting prior explanations from similar training examples.

The abductive framework is empirically evaluated on commonsense and scientific question answering. In particular, the experiments demonstrate that the proposed framework can be instantiated with sparse and dense pre-trained encoders to address multi-hop inference without direct supervision, or adopted as evidence retrievers for downstream Transformers, achieving strong performance when compared to existing explanation-based approaches. Moreover, the chapter investigates the impact of the *retrieve-reuse-refine* paradigm on semantic drift, showing that it boosts the quality of the most challenging explanations, resulting in improved robustness and accuracy in downstream inference tasks[1].

4.1 Introduction

Multi-hop inference is the task of composing two or more pieces of evidence from external knowledge resources to address a particular reasoning problem [151]. In the context of Natural Language Inference (NLI), this task is often used to develop

[1]This chapter follows the publication "Case-based Abductive Natural Language Inference"[160]

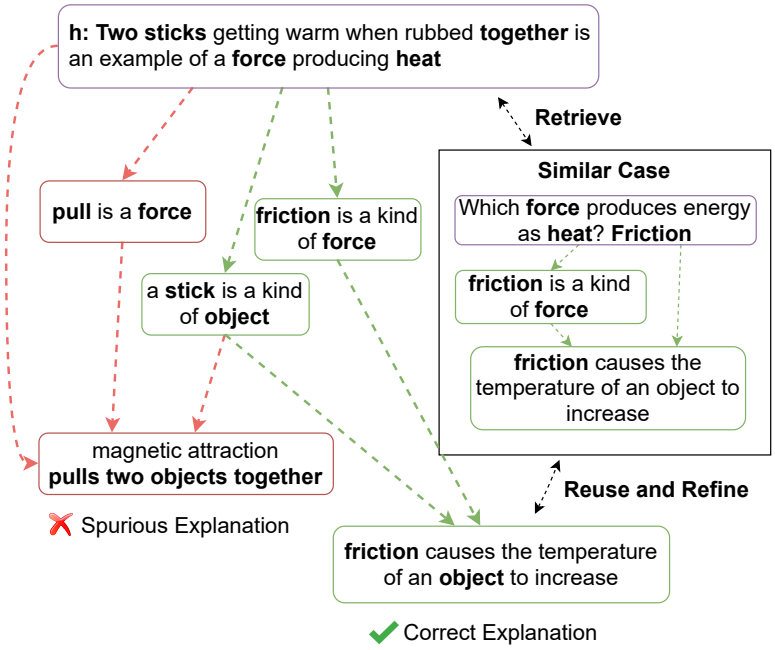


Figure 4.1: Performing multi-hop inference considering each case in isolation can lead to the construction of spurious explanations. In contrast, we propose the adoption of the *retrieve-reuse-refine* paradigm in *case-based reasoning*.

and evaluate explanation-based systems, capable of performing transparent multi-step reasoning with natural language [173].

While multi-hop inference has been largely explored for *extractive* problems such as open-domain question answering [187], increasing attention is being dedicated to the *abstractive* setting, where the models are required to compose long chains of facts expressing abstract commonsense and scientific knowledge [72, 22]. In this setting, multi-hop inference is often framed as an abductive natural language inference problem, where, for a given set of alternative hypotheses $H = \{h_1, h_2, \ldots, h_n\}$, the goal is to construct an explanation for each $h_i \in H$ and select the hypothesis supported by the *best explanation*. Existing approaches address abductive inference considering each test hypothesis in isolation, employing iterative and path-based methods [94, 185] or explicit constraints to guide the generation of a plausible explanation graph supporting the correct answer [84, 89].

However, this paradigm poses several challenges in the abstractive setting as: (1) the structure of the explanation is not evident from the decomposition of the hypothesis, that is, the type of facts required for the inference cannot be derived from the surface form of the reasoning problem; (2) core explanatory facts tend to be abstract, sharing a low number of terms with the hypothesis, making it hard to correctly estimate their

relevance for the inference; (3) background knowledge sources contain a large amount of distracting information overlapping with the hypothesis, which can lead to the generation of spurious explanations. Consequently, existing approaches often suffer from a phenomenon known as *semantic drift* [82] – i.e., the tendency of composing incorrect reasoning chains leading to wrong conclusions. The example in Figure 4.1 illustrates some of these challenges.

In contrast with the dominant paradigm, we propose to integrate Abductive NLI in a *case-based reasoning* framework [134]. Case-based reasoning systems operate on the hypothesis that similar problems require similar solutions, addressing new cases via analogical transfer from previous cases solved in the past. Specifically, the case-based reasoning framework employs a *retrieve-reuse-refine* paradigm to model inference over unseen problems [135, 36]. In the context of multi-hop natural language inference, we hypothesise that the adoption of a case-based reasoning framework can help tackle some of the challenges involved in the abstractive setting since: (1) similar hypotheses tend to require similar explanations; (2) abstract facts tend to express general explanatory knowledge about underlying regularities, being frequently reused to explain a large variety of hypotheses; (3) prior solutions can explicitly help constrain the search space for new problems, reducing the risk of composing spurious inference chains. To this end, we present a case-based abductive NLI model that retrieves and adapts natural language explanations from training examples to construct new explanations for unseen cases and address downstream NLI problems through explanation-based inference.

Specifically, this chapter provides the following contributions: (1) To the best of our knowledge, we are the first to propose an end-to-end case-based abductive framework for multi-hop NLI; (2) We empirically demonstrate the efficacy of the case-based framework on commonsense and scientific question answering, showing that the proposed model can be effectively integrated with different sentence encoders and downstream Transformers, achieving strong performance when compared to existing multi-hop and explanation-based approaches; (3) We study the impact of the retrieve-reuse-refine paradigm on semantic drift, and how this affects accuracy and robustness in downstream inference. Our results show that the case-based framework boosts the quality of the explanations for the most challenging hypotheses, resulting in improved accuracy in downstream question answering.

4.2 Case-based Abductive NLI

To investigate ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"* for Abductive NLI, we present a case-based reasoning framework that is guided by the following research hypotheses:

- **RH3.2**: Similar NLI problems tend to require similar explanatory patterns, sharing abstract explanatory statements expressing general underlying regularities;
- **RH3.3**: Prior solutions from similar cases can explicitly help constrain the search space for new cases, reducing the risk of composing spurious inference chains and improving accuracy on downstream Abductive NLI.

For a given set of alternative natural language hypotheses $H = \{h_1, h_2, \ldots, h_n\}$, the goal is to construct an explanation for each $h_i \in H$ and select the hypothesis supported by the *best explanation*. Given an hypothesis h_i (e.g., *"Two sticks getting warm when rubbed together is an example of a force producing heat"*), we construct an explanation justifying h_i by extracting and composing inference chains between multiple explanatory facts retrieved from an external corpus.

To generate an explanation for h_i, we adopt a *case-based reasoning* paradigm composed of three major phases, *retrieve-reuse-refine*, which can be summarised as follows:

1. **Retrieve:** In the retrieve phase, we employ a sentence encoding mechanism to search over two distinct embedding spaces. A first embedding space (*Facts Embeddings*) is adopted to retrieve a set of candidate explanatory sentences for the hypothesis. A second embedding space (*Cases Embeddings*) is used to retrieve similar cases solved in the past whose explanations can be useful to guide the search for a new solution.
2. **Reuse:** In the reuse phase, we condition the relevance of a given fact on the set of explanations retrieved from the most similar cases. Specifically, we reuse previously solved cases to estimate the *explanatory power* of a fact, representing the extent to which a given fact is used in explanations for past hypotheses.
3. **Refine:** In this phase, the list of candidate explanatory facts is refined to build the final explanation. We model the construction of an explanation via multi-hop inference between hypothesis and candidate facts, composing abstractive

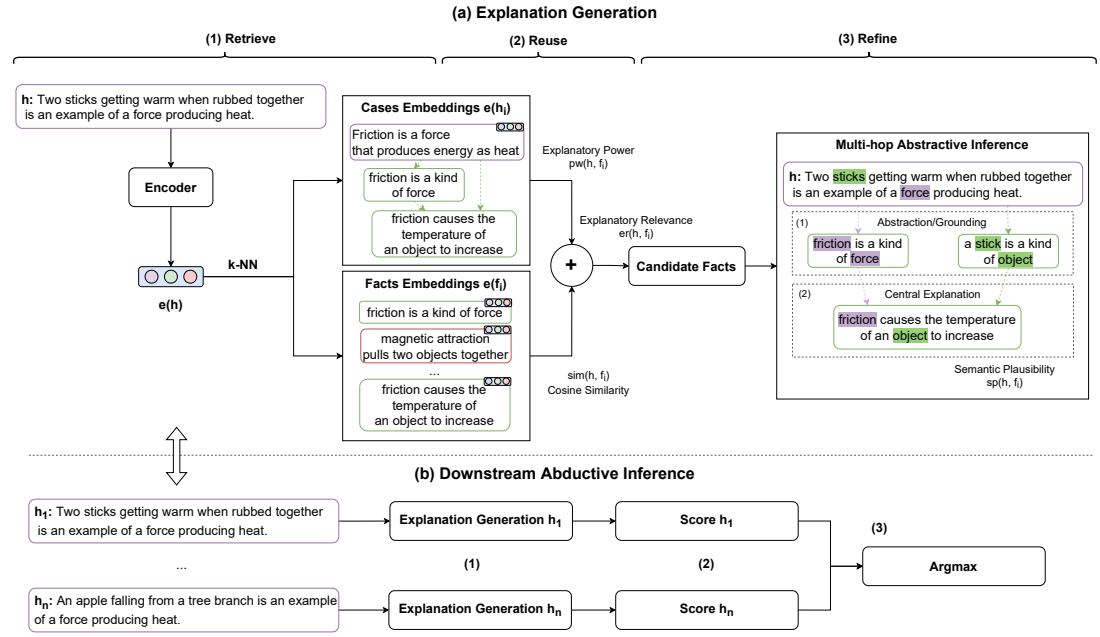


Figure 4.2: Overview of the proposed framework. We adopt a *retrieve-reuse-refine* paradigm to construct explanations for unseen hypotheses (a) and address downstream NLI tasks via explanation-based inference (b).

inference chains to estimante the plausibility of central explanatory sentences.

Given a set of alternative hypotheses, we adopt the case-based reasoning framework for explanation generation, and subsequently leverage the score assigned to each explanation to address downstream inference tasks. Additional details on the *retrieve-reuse-refine* phases are described in the following sections.

4.3 Explanation Generation

4.3.1 Retrieve

We perform k-NN search over two distinct embedding spaces: (a) an embedding space encoding individual commonsense and scientific facts that can be used to construct new explanations (*Facts Embeddings*); (b) an embedding space of true hypotheses associated with their respective explanations (*Cases Embeddings*). An explanation for a given hypothesis h_i is a composition of facts, $E_i = \{f_1, \ldots, f_n\}$.

To perform k-NN search, we employ a sentence encoder $e(\cdot)$. Specifically, we use $e(\cdot)$ to derive a vector for the test hypothesis h and adopt cosine similarity to efficiently

score and rank facts and hypotheses in the embedding spaces, retrieving the top-k instances. We perform our experiments using a sparse (BM25 [129]) and a pre-trained dense encoder (Sentence-BERT [125]) adopting a search index for efficient retrieval (IndexIVFFlat in FAISS [76], subsequently calling it "Faiss Index" – cf. Appendix C.1.2). We adopt the WorldTree corpus [72] as background knowledge.

4.3.2 Reuse

Chapter 2 and 3 have shown that explanatory facts expressing underlying regularities tend to create explanatory patterns across similar hypotheses. Following these results, we conjecture that explanations from similar cases can be used to constraining the search space for unseen hypotheses at inference time.

Specifically, following Chapter 3, given an unseen hypothesis h and a fact f_i, we adopt the explanations retrieved from the top-K similar hypotheses in the *Case Embeddings* to estimate the *explanatory power* of f_i:

$$pw(h, f_i) = \sum_{h_k \in kNN(h)}^{K} sim(e(h), e(h_k)) \cdot \mathbb{1}(f_i, h_k) \tag{4.1}$$

$$\mathbb{1}(f_i, h_k) = \begin{cases} 1 & \text{if } f_i \in E_k \\ 0 & \text{if } f_i \notin E_k \end{cases} \tag{4.2}$$

where $kNN(h) = \{h_1, \ldots, h_K\}$ represents the list of k-nearest hypotheses of h retrieved according to the cosine similarity $sim(\cdot)$ between the embeddings $e(h)$ and $e(h_k)$, and $\mathbb{1}(\cdot)$ is the indicator function verifying if f_i is included in the explanation E_k for the hypothesis h_k. Therefore, for each hypothesis h_k in the set of k-nearest neighbours, the model sums up the quantity $sim(\cdot)$ only if f_i is used to explain h_k. Since $sim(e(h), e(h_k))$ represents the similarity between h and h_k, the more f_i explains past hypotheses that are similar to h the higher the explanatory power of f_i. To condition the list of candidate explanatory facts on previously solved cases while controlling for relevance with respect to the test hypothesis h, we compute the final *explanatory relevance* of each f_i by interpolating the explanatory power with the similarity between the embeddings $e(h)$ and $e(f_i)$:

$$er(h, f_i) = \lambda \cdot sim(e(h), e(f_i)) + (1 - \lambda) \cdot pw(h, f_i) \tag{4.3}$$

The explanatory relevance score is used to re-rank and filter the list of candidate facts for the subsequent phase.

4.3.3 Refine

In the refine phase, the model considers the set of candidate facts retrieved in the previous stage to construct the final explanation for h. We model the construction of an explanation through multi-hop inference between hypothesis and candidate facts via the composition of explicit inference chains. To this end, we represent facts and hypothesis as sets of distinct concepts $CP(s_i) = \{cp_1, \ldots, cp_n\}$ (e.g., *"friction is a kind of force"* is represented as the set $\{friction, force\}$, details in the appendix), and connect two generic sentences s_i and s_j by means of shared concept in $CP(s_i) \cap CP(s_j)$.

Following the corpus analysis performed in Chapter 2, we model multi-hop inference as an explicit abstraction process, attempting to estimate the plausibility of abstract explanatory sentences. Specifically, we model abstraction through the construction of an explanation graph in different stages, starting with the hypothesis h as the only node. In the first stage, the model extends the graph with the facts that share direct concepts with h and that express taxonomic relations or synonymy. This step can be seen as an abstraction/grounding mechanism aimed at linking the hypothesis to core explanatory statements with high explanatory power (e.g., linking *stick* to *object* and *friction* to *force* in Figure 4.2).

In the second stage, the model extends the graph with the remaining candidate explanatory facts that share at least one concept with previously added nodes. We consider these facts as the *central explanatory nodes*. After constructing the graph, we leverage its structure to estimate the *semantic plausibility* of the sentences f_i corresponding to the central explanatory nodes:

$$sp(h, f_i) = \frac{\sum_{cp_j \in CP(h)} path(cp_j, f_i)}{|CP(h)|} \tag{4.4}$$

where $path(cp_j, f_i)$ is equal to 1 if there exists a path in the graph connecting the concept cp_j in the hypothesis with some concepts in the fact f_i, 0 otherwise. Therefore, the semantic plausibility represents the percentage of concepts in the hypothesis h that have a path in the graph leading to f_i.

To derive the final explanation while conditioning on previously solved cases, we sum the *explanatory relevance* computed during the reuse phase with the *semantic plausibility*, pruning the graph considering only the top n central explanatory sentences and their linked grounding nodes (Fig. 4.2).

4.4 Abductive Inference

Given a set of alternative hypotheses $H = \{h_1, \ldots, h_n\}$, we adopt the model for abductive inference by generating an explanation for each hypothesis and selecting as an answer the one supported by the best explanation. To this end, we assign a score to each hypothesis h_i in H equal to the sum of the scores of the central facts included in the explanation for h_i.

4.5 Empirical Evaluation

Experimental Setup. We evaluate the Case-based Abductive NLI (CB-ANLI) framework on WorldTree [72] and AI2 Reasoning Challenge (ARC) [22], two multiple-choice science question answering datasets designed to test abstractive commonsense and scientific inference. To perform the experiments, we transform each question-candidate answer pair into a hypothesis following the methodology described in [37].

The knowledge bases required for the inference are populated using the WorldTree corpus [72]. The corpus contains a large set of commonsense and scientific facts ($\approx 10K$) that are used to construct explanations for multiple-choice science questions. The explanations include an average of 6 facts (and as many as $\approx$ 20), requiring challenging multi-hop inference to be generated. We store the individual facts for deriving the *FactsEmbeddings* and consider the *training* questions ($\approx$ 1K) and their explanations as the set of previously solved cases (*CasesEmbeddings*). For the refine phase, we dynamically extract the concepts in facts and hypotheses using WordNet with NLTK[2]. Additional details are described in the appendix.

Sentence Encoders. We evaluate CB-ANLI using sparse and dense sentence encoders without additional training. The sparse version adopts BM25 vectors [129], while the dense version employs Sentence-BERT (large) [125, 150].

4.5.1 WorldTree

In this section, we present the results achieved on the WorldTree test-set (1247 questions). We report the accuracy of the case-based framework with different numbers n of central facts in the explanations. We compare the proposed framework against different

[2] https://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html

Model	Overall	Easy	Challenge
Sparse Retrieval Solver			
BM25 (k = 1) [22]	41.21	44.96	32.99
BM25 (k = 2)	43.62	48.54	32.73
BM25 (k = 3)	45.87	50.76	35.05
Dense Retrieval Solver			
S-BERT (k = 1) [125]	44.91	50.99	31.44
S-BERT (k = 2)	45.79	51.45	33.25
S-BERT (k = 3)	44.51	49.82	32.73
Path-based Solver			
PathNet [94]	41.50	43.32	**36.42**
Transformers			
BERT-large [39]	46.19	52.62	31.96
RoBERTa-large [108]	**50.20**	**57.04**	35.05
Case-based Abductive NLI			
CB-ANLI BM25 (n = 1)	52.13	56.34	42.78
CB-ANLI BM25 (n = 2)	**55.17**	60.42	**43.56**
CB-ANLI BM25 (n = 3)	52.69	58.56	39.69
CB-ANLI S-BERT (n = 1)	54.45	**61.23**	39.43
CB-ANLI S-BERT (n = 2)	52.77	59.60	37.62
CB-ANLI S-BERT (n = 3)	51.64	58.67	36.08

Table 4.1: Accuracy on WorldTree (test-set) for *easy* and *challenge* questions.

categories of approaches: *Retrieval Solvers*, *Path-based Solvers*, and *Transformers*. The results in terms of question answering accuracy are reported in Table 4.1.

Retrieval Solvers. We employ stand-alone BM25 and Sentence-BERT (large) as sparse and dense retrieval solvers [22]. Given an hypothesis h, the solvers retrieve the most k relevant facts for h using cosine similarity. The cosine similarity scores are then summed up to determine the best hypothesis. These baselines use the same encoders adopted by our model. However, we observe that CB-ANLI is able to outperform both sparse and dense retrieval models by up to $\approx 10\%$ accuracy, demonstrating the decisive role of the proposed case-based paradigm.

Path-based Solvers. We consider PathNet [94] as a multi-hop inference baseline. This model constructs inference paths connecting question and candidate answer, and

RoBERTa + Retriever	Over.	Easy	Chal.
BM25 (k = 1)	57.06	60.88	48.57
BM25 (k = 2)	61.07	66.82	48.32
BM25 (k = 3)	61.23	65.54	51.12
S-BERT (k = 1)	55.85	61.46	43.41
S-BERT (k = 2)	60.91	66.82	47.80
S-BERT (k = 3)	56.96	62.04	45.73
CB-ANLI BM25 (n = 1)	61.71	66.82	50.38
CB-ANLI BM25 (n = 2)	**63.48**	**69.38**	50.38
CB-ANLI BM25 (n = 3)	62.43	67.77	50.63
CB-ANLI S-BERT (n = 1)	59.99	65.54	47.45
CB-ANLI S-BERT (n = 2)	63.32	67.98	**52.97**
CB-ANLI S-BERT (n = 3)	62.27	67.63	50.38

Table 4.2: Accuracy of RoBERTa large fine-tuned on the WorldTree test-set and augmented with different explanation models.

subsequently scores them through a neural encoder to derive the correct answer. We reproduce PathNet using the source code available online[3]. Contrary to CB-ANLI, PathNet does not adopt a case-based reasoning framework to construct the explanations, considering each test hypothesis in isolation. We observe that CB-ANLI can significantly outperform PathNet with up to $\approx 13\%$ improvement overall and $\approx 7\%$ on challenge questions.

Transformers. We compare CB-ANLI against BERT large [39] and RoBERTa large [108] fine-tuned on the multiple-choice question answering task. We observe that the proposed approach is able to outperform both RoBERTa and BERT (up to $\approx 5\%$ and $\approx 9\%$ respectively).

4.5.2 Transformers with Explanations

We evaluate CB-ANLI as an evidence retrieval model by combining it with downstream Transformers. To perform this experiment, we augment the input of RoBERTa large with the explanations constructed for each hypothesis, and fine-tune the model to maximise the score for the correct one. Table 4.2 reports the accuracy achieved with RoBERTa large when adopting CB-ANLI and stand-alone models as evidence retrievers. We observe that RoBERTa augmented with CB-ANLI achieves better overall results

[3] https://github.com/allenai/PathNet

Previous Explainable Models	**Accuracy**
TupleInf [89]	23.83
TableILP [84]	26.97
DGEM [22]	27.11
KG^2 [188]	31.70
Unsupervised AHE [184]	33.87
Supervised AHE [184]	34.47
ET-RR [117]	36.61
ExplanationLP [152]	40.21
AutoROCC [185]	41.24
Case-based Abductive NLI	
CB-ANLI BM25 (n = 1)	33.45
CB-ANLI BM25 (n = 2)	34.39
CB-ANLI BM25 (n = 3)	33.79
CB-ANLI S-BERT (n = 1)	**36.77**
CB-ANLI S-BERT (n = 2)	35.75
CB-ANLI S-BERT (n = 3)	34.30
CB-ANLI S-BERT (n = 1) + RoBERTa	44.02
CB-ANLI S-BERT (n = 2) + RoBERTa	**47.86**
CB-ANLI S-BERT (n = 3) + RoBERTa	42.40

Table 4.3: Performance on the AI2 Reasoning Challenge (ARC)

for each value of n, suggesting that the proposed framework is able to generate more discriminating explanations for downstream language models.

4.5.3 ARC Challenge

To evaluate the generalisation of CB-ANLI on a broader set of challenge questions, we run additional experiments on the AI2 Reasoning Challenge (ARC) [22]. Here, we keep the same configuration and set of hyperparameters. Table 4.3 reports the results achieved on the test-set (1172 challenge questions). We observe that CB-ANLI with Sentence-BERT can generalise better on ARC. We attribute these results to the ability of Sentence-BERT to go beyond lexical overlaps for case retrieval, supporting generalisation on new hypotheses with different surface forms. To show the impact of evidence retrieval on ARC, we fine-tune RoBERTa with the explanations constructed by the Sentence-BERT version. For a fair comparison, we compare CB-ANLI against published explanation-based approaches that are fine-tuned only on ARC, without additional pre-training on related datasets (e.g. OpenBookQA [111], RACE [96]). The

Paradigm	Overall	Easy	Challenge
CB-ANLI BM25			
Retrieve-Reuse-Refine	**55.17**	**60.42**	**43.56**
Retrieve-Reuse	49.00	55.18	35.30
Retrieve-Refine	43.46	46.57	36.60
CB-ANLI S-BERT			
Retrieve-Reuse-Refine	**54.45**	**61.23**	**39.43**
Retrieve-Reuse	47.79	53.55	35.05
Retrieve-Refine	42.66	47.48	32.21

Table 4.4: Ablation Study on WorldTree (test-set).

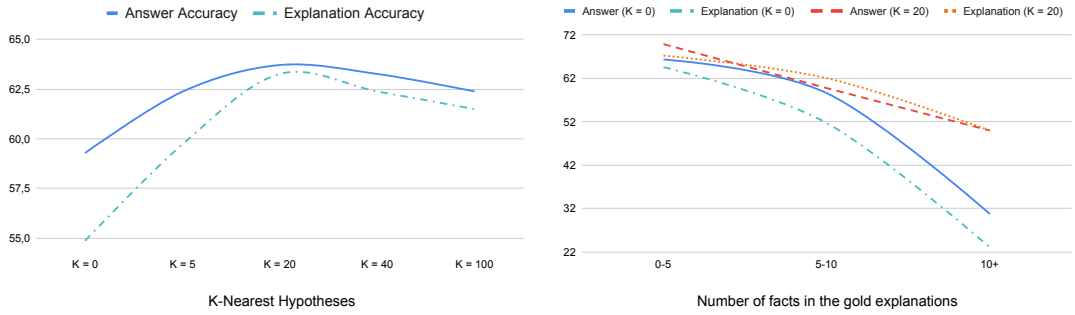
results show that CB-ANLI (Sentence-BERT) is third in the ranking, outperforming explainable systems based on Integer Linear Programming (ILP) [89, 84] and pre-trained embeddings [184]. At the same time, CB-ANLI obtains competitive results when compared with most of the fine-tuned neural approaches, including ET-RR [117]. Moreover, when combined with RoBERTa, CB-ANLI achieves the best results among the considered approaches, improving on AutoROCC [185] by $\approx 6\%$.

4.5.4 Ablation Study

We carried out an ablation study to investigate the impact of the case-base reasoning framework on downstream inference performance. To this end, we consider different versions of CB-ANLI by alternatively removing the impact of the reuse and refine phase. For the first, we remove the *explanatory power* term in Equation 4.3. For the latter, we simply skip the refine phase ignoring the *semantic plausibility* to filter the central explanatory facts. The results of the study, reported in Table 4.4, demonstrate the key role of each phase to achieve the final inference performance.

4.5.5 Impact on Semantic Drift

In this section, we investigate the impact of the *retrive-reuse-refine* paradigm on semantic drift, and how this affects the results on downstream reasoning tasks and explanation quality. To this end, we measure the performance of CB-ANLI when considering a different number K of previously solved hypotheses (when $K = 0$ the model is equivalent to a non-case-based method).



(a) (b)

Figure 4.3: Impact of the case-based framework on semantic drift. K represents the number of similar hypotheses considered for computing the explanatory power (Worldree dev-set).

To evaluate the quality of the generated explanations, we use the annotated explanations in the WorldTree corpus as gold standards, computing the accuracy of the explanations constructed by the model as the percentage of the selected central facts that are part of the gold explanations. Since the explanations in the *test-set* are not publicly available, we perform this analysis on the *dev-set*.

Figure 4.3 (a) illustrates the change in answer and explanation accuracy on WorldTree with an increasing number K of similar cases. The graph demonstrates that the improvement in answer prediction is associated with better explanation generation (with a peak at $K = 20$). Specifically, by conditioning the inference on an increasing number of similar hypotheses, CB-ANLI is able to construct more accurate explanations, a feature that has a direct impact on downstream inference performance.

Figure 4.3 (b) shows the accuracy of the model on hypotheses requiring longer explanations when compared to a non-case-based version ($K = 0$). In general, a higher number of facts in the gold explanation is associated with a higher probability of semantic drift [71]. The graph confirms a strong relation between explanation accuracy and question answering accuracy, and demonstrates that the improvement obtained through the case-based framework is particularly evident on the most challenging inference problems (10+ facts in the explanations). This results allow us to conclude that the case-based reasoning framework has a key role in alleviating semantic drift during multi-hop inference.

Test Question	Prediction	Constructed Explanation ($K = 20, n = 1$)	Accurate
What force is needed to help stop a child from slipping on ice? (A) gravity, (B) friction, (C) electric, (D) magnetic	(B) friction	(1) counter means reduce; stop; resist; (2) ice is a kind of object; (3) slipping is a kind of motion; (4) stop means not move; **(5) friction acts to counter the motion of two objects when their surfaces are touching**	**Y**
What causes a change in the speed of a moving object? (A) force, (B) temperature, (C) change in mass (D) change in location	(A) force	**(1) a force continually acting on an object in the same direction that the object is moving can cause that object's speed to increase in a forward motion**	N
Weather patterns sometimes result in drought. Which activity would be most negatively affected during a drought year? (A) boating, (B) farming, (C) hiking, (D) hunting	(B) farming	(1) affected means changed; (2) a drought is a kind of slow environmental change; **(3) farming changes the environment**	N
Beryl finds a rock and wants to know what kind it is. Which piece of information about the rock will best help her to identify it? (A) The size of the rock, (B) The weight of the rock, (C) The temperature where the rock was found, (D) The minerals the rock contains	(A) The size of the rock	(1) a property is a kind of information; (2) size is a kind of property; (3) knowing the properties of something means knowing information about that something. **(4) the properties of something can be used to identify; used to describe that something**	**Y**
Jeannie put her soccer ball on the ground on the side of a hill. What force acted on the soccer ball to make it roll down the hill? (A) gravity, (B) electricity, (C) friction, (D) magnetism	(C) friction	(1) the ground means Earth's surface; (2) rolling is a kind of motion; (3) a roll is a kind of movement; **(4) friction acts to counter the motion of two objects when their surfaces are touching**	N

Table 4.5: Examples of explanations constructed for the predicted answers. The underlined choices represent the correct answers. *Accurate* indicates whether the central fact **(bold)** is labelled as a gold explanation in the corpus.

4.5.6 Faithfulness and Error Analysis

Finally, we present an analysis on the faithfulness of the model, investigating the relation between correct/wrong answer prediction and accurate/inaccurate explanations. Overall, we found that a total of 81.25% of the correct answers are derived from accurate explanations. This situation is illustrated in the first example in Table 4.5. On the other hand, a total of 18.75% of correct answers are derived from inaccurate explanations (second and third rows in the table). However, as shown in the second example, we observe that CB-ANLI can sometimes find alternative ways of constructing plausible explanations, considered inaccurate only because of a mismatch with the corpus annotation. In contrast, the example number 4 shows the case in which an accurate explanation is not sufficient to discriminate the correct answer (this case occurring for a total of 31.71% of incorrect answers). Finally, the last row describes the situation in which wrong answers are caused by inaccurate explanations (68.29% of the time). This analysis demonstrates the interpretability of the framework, showing that its behaviour can be generally traced back to the quality of the generated explanations.

4.6 Related Work

Performing multi-hop inference for abstractive NLI tasks is challenging as the general structure of the explanations cannot be derived from the surface form of the problem. Previous work has demonstrated that models in this setting are affected by semantic drift – i.e., the construction of spurious explanations leading to wrong conclusions [46, 82]. Existing approaches frame multi-hop inference as the problem of building an optimal graph, conditioned on a set of semantic constraints [85, 89, 68, 84], or adopting iterative methods, using sparse or dense encoding mechanisms [184, 185, 122, 94]. Our model is related to previous work that leverages annotated explanations to reduce semantic drift [180, 72]. However, this work is limited to explanation regeneration tasks [71, 18, 34], and their impact and applicability on downstream NLI has yet to be explored. In this chapter, we move a step forward, exploring the impact of annotated explanations on semantic drift for end-to-end inference problems.

Case-based Reasoning. Our approach is related to previous work on case-based reasoning [134, 135, 36]. Similar to the retrieve-reuse-refine paradigm adopted in case-based reasoning, we employ encoding mechanisms to retrieve explanations for cases solved in the past, and adapt them in the solution of new problems. Recent work in NLP investigates the use of a similar paradigm via k-NN retrieval on training examples. [81, 80] adopt k-NN search to retrieve similar training examples and improve pre-trained language models and machine translation without additional training. Similarly, [35, 33] propose a case-based framework for knowledge base reasoning, while [78] reuse similar cases to improve BERT [39] on cloze-style QA. To the best of our knowledge, this is the first application of case-based reasoning for multi-hop inference on commonsense and scientific NLI tasks.

Neuro-symbolic Models. The work presented in this chapter is related to hybrid neuro-symbolic approaches for multi-hop inference. In this context, most of the existing approaches combine neural models with symbolic programs [107, 75, 20, 42, 183, 170]. For instance, [75] propose the adoption of a Neural Module Network [3] for multi-hop question answering by designing four atomic neural modules (Find, Relocate, Compare, NoOp). Similarly, [170] propose a methodology to perform multi-hop inference using Prolog via the integration of a question decomposition model and a weak unification mechanism. However, differently from the methodology discussed in this chapter, these approaches have been generally applied to extractive tasks, assuming that the

structure of the inference (and, therefore, the explanation) can be derived from a direct decomposition of the NLI problem.

4.7 Conclusion

This chapter presented CB-ANLI, a model that integrates multi-hop and case-based reasoning in a unified framework. We demonstrated the efficacy of reusing explicit explanatory patterns for complex abstractive NLI tasks. In particular, the chapter demonstrated the impact of the case-based framework on commonsense and scientific question answering, showing that the proposed model can be effectively integrated with different sentence encoders and downstream Transformers, achieving strong performance when compared to existing multi-hop and explainable approaches. Moreover, the chapter investigated the impact of the retrieve-reuse-refine paradigm on semantic drift, and how this affects accuracy and robustness for downstream inference. Our results show that the case-based framework boosts the quality of the explanations for the most challenging hypotheses, resulting in improved accuracy in downstream question answering.

4.8 Scoping and Limitations

The explanatory power model adopted in the reuse phase inherit the limitations discussed in Chapter 3. In the current implementation of CB-ANLI, the refine phase adopts some simplified assumptions to model the abstraction process required for explanation generation. This process, in fact, is performed by assuming that abstraction at the concept level translates to a correct mapping between hypotheses and central explanatory sentences. However, contextual linguistic elements should be taken into account in this process as they can affect the overall meaning of the sentence and of the specific concept being abstracted. While contextual element are partially considered during the precedent phases and have been shown to improve robustness, additional work is required to guarantee the correctness of the refine phase and the adopted abstractive mechanism.

Chapter 5

Hybrid Autoregressive Inference

This chapter aims to investigate ***RQ4:*** *"Can hybrid models integrating latent and explicit representations provide a framework for a better accuracy-scalability trade-off in Explanation-based NLI?"*. To this end, the chapter focuses on scalable bi-encoder architectures, investigating the problem of scientific Explanation Regeneration at the intersection of dense and sparse models. Specifically, the chapter presents **SCAR** (for Scalable Autoregressive Inference), a hybrid explanation-based inference framework that iteratively combines a Transformer-based bi-encoder with the sparse model of explanatory power discussed in Chapter 3.

The experiments demonstrate that the hybrid framework significantly outperforms previous sparse models, achieving performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster and scalable to corpora of millions of facts. Further analyses on semantic drift reveal that the proposed hybridisation boosts the performance in multi-hop inference, contributing to improved accuracy when addressing explanation-based question answering in an iterative fashion[1].

5.1 Introduction

Explanation Regeneration is the task of retrieving and combining two or more facts from an external knowledge source to reconstruct the evidence supporting a certain natural language hypothesis [180, 72]. As such, this task represents a crucial intermediate step for the development and evaluation of Explanation-based NLI models [173, 151]. In particular, Explanation Regeneration on science questions has been proposed as a

[1]This chapter follows the publication "Hybrid Autoregressive Inference for Scalable Multi-hop Explanation Regeneration"[159]

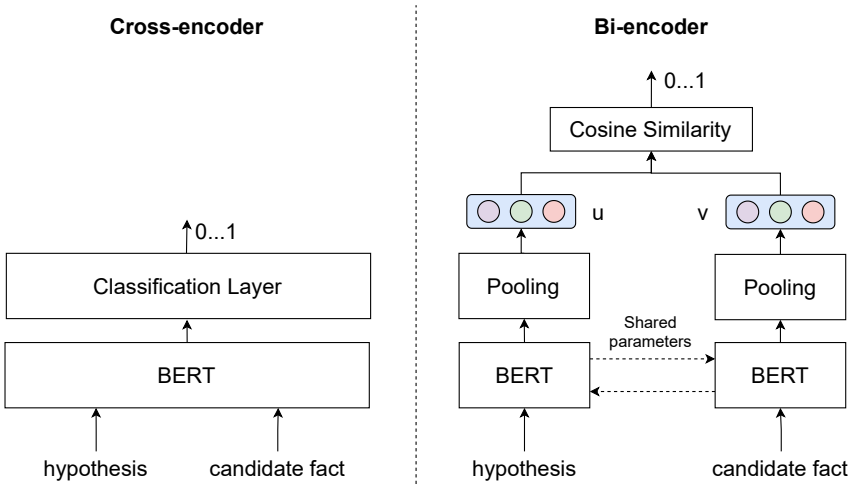


Figure 5.1: Overview of the cross-encoder and bi-encoder architecture. Cross-encoders (left) tend to be more robust thanks to the application of a classification mechanism. However, in contrast to bi-encoders (right), the representation vectors of candidate facts cannot be pre-computed and stored in apposite dense indexes for efficient inference [125].

benchmark for complex multi-hop and explanation-based inference [71]. Scientific explanations, in fact, require the articulation and integration of commonsense and scientific knowledge for the construction of long explanatory reasoning chains, making multi-hop inference particularly challenging for existing models [22, 86]. Moreover, since the structure of scientific explanations cannot be derived from the decomposition of the questions, the task requires the encoding of abstraction and grounding mechanisms for the identification of relevant explanatory knowledge [152, 67].

To tackle these challenges, existing neural approaches leverage the power of the self-attention mechanism in Transformers [39, 163], training sequence classification models (i.e., cross-encoders) on annotated explanations to compose relevant explanatory chains [19, 34, 21, 6]. While Transformers achieve state-of-the-art performance, cross-encoders make multi-hop inference intrinsically inefficient and not scalable to large corpora. The cross-encoder architecture, in fact, does not allow for the construction of dense indexes to cache the encoded explanatory sentences, resulting in prohibitively slow inference time for real-world applications [63].

In this chapter, we are interested in developing new mechanisms to enable scientific Explanation Regeneration at scale, optimising, at the same time, quality of the explanations and inference time. To this end, we focus our attention on bi-encoders (or

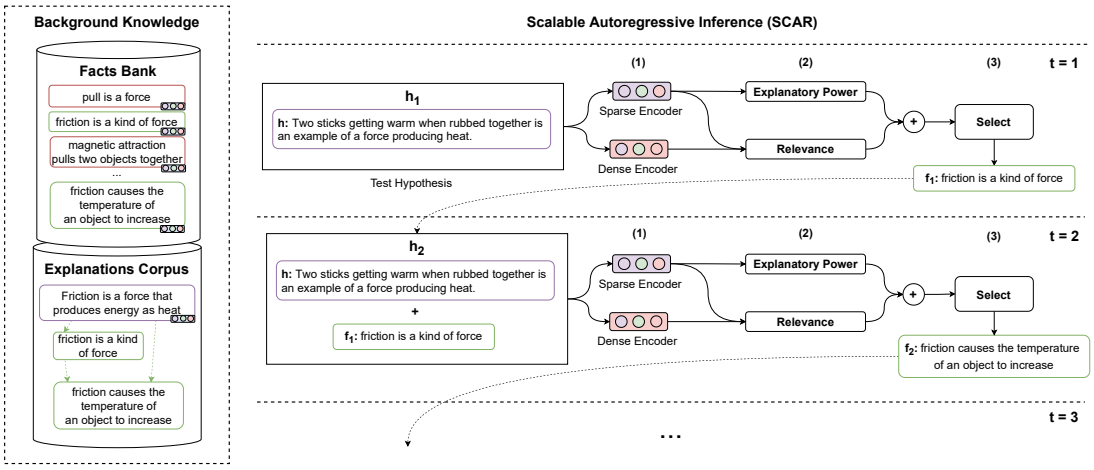


Figure 5.2: We propose a hybrid, scalable Explanation Regeneration model that performs inference autoregressively. At each time-step t, we perform inference integrating sparse and dense bi-encoders (1) to compute relevance and explanatory power of sentences in the fact bank (2) and expand the explanation (3). The relevance of a fact at time-step t is conditioned on the partial explanation constructed at time $t-1$, while the explanatory power is estimated leveraging inference patterns emerging across similar hypotheses in the Explanations Corpus.

siamese networks) [126], which allow for efficient inference via Maximum Inner Product Search (MIPS) [76]. Given the complexity of multi-hop reasoning in the scientific domain, bi-encoders are expected to suffer from a drastic drop in performance since the self-attention mechanism cannot be leveraged to learn meaningful compositions of explanatory chains. However, we hypothesise that the orchestration of latent and explicit patterns emerging in natural language explanations can improve the quality of the inference while preserving the scalability intrinsic in bi-encoders.

To validate this hypothesis, we present **SCAR** (for Scalable Autoregressive Inference), a hybrid architecture that combines a Transformer-based bi-encoder with a sparse model of explanatory power, designed to capture explicit inference patterns in corpora of scientific explanations. Specifically, SCAR integrates sparse and dense encoders to define a joint model of relevance and explanatory power and perform inference in an iterative fashion, conditioning the probability of selecting a fact at time-step t on the partial explanation constructed at time-step $t-1$ (Fig. 5.2). We performed an extensive evaluation on the WorldTree corpus [71], presenting the following conclusions:

1. The hybrid framework based on bi-encoders significantly outperforms existing sparse models, achieving performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster.

2. We study the impact of the hybridisation on semantic drift, showing that it makes SCAR more robust in the construction of challenging explanations requiring long reasoning chains.

3. We investigate the applicability of SCAR on multi-hop question answering without additional training, demonstrating improved accuracy and robustness when performing explanation-based inference iteratively.

4. We perform a scalability analysis by gradually expanding the adopted fact bank, showing that SCAR can scale to corpora containing millions of facts.

To the best of our knowledge, we are the first to propose a hybrid autoregressive model for complex multi-hop inference in the scientific domain, demonstrating its efficacy for Explanation Regeneration at scale.

5.2 Multi-hop Explanation Regeneration

Given a scientific hypothesis *h* expressed in natural language (e.g., *"Two sticks getting warm when rubbed together is an example of a force producing heat"*), the task of Explanation Regeneration consists in reconstructing the evidence supporting *h*, composing a sequence of atomic sentences $E_{seq} = f_1, \ldots, f_n$ from external corpora (e.g., f_1*:"friction is a kind of force";* f_2*:"friction causes the temperature of an object to increase"*). Explanation Regeneration can be framed as a multi-hop abductive inference problem, where the goal is to construct the best explanation supporting a given natural language statement adopting multiple retrieval steps.

To learn to regenerate scientific explanations, a recent line of research relies on explanation-centred corpora such as WorldTree [71], which are typically composed of two distinct knowledge sources (Fig. 5.2):

1. A *fact bank* of individual commonsense and scientific sentences including the knowledge necessary to construct explanations for scientific hypotheses.

2. An *explanations corpus* consisting of true hypotheses and natural language explanations composed of sentences from the fact bank.

5.3 Hybrid Autoregressive Inference

To investigate ***RQ4:*** *"Can hybrid models integrating latent and explicit representations provide a framework for a better accuracy-scalability trade-off in Explanation-based NLI?"*, we present a hybrid framework that is guided by the following research hypotheses:

- **RH4.1**: Sparse and dense representations possess complementary features for modelling the notion of explanatory relevance;
- **RH4.2**: Dense representations based on bi-encoders can be effectively integrated with Explanation Regeneration models that leverage explicit explanatory patterns, improving accuracy in multi-hop inference while preserving scalability.

To model the multi-hop nature of scientific explanations, we propose a hybrid architecture that performs inference autoregressively (Fig. 5.2). Specifically, we model the probability of composing an explanation sequence $E_{seq} = f_1, \ldots, f_n$ for a certain hypothesis h using the following formulation:

$$P(E_{seq}|h) = \prod_{t=1}^{n} P(f_t|h, f_1, \ldots f_{t-1}) \tag{5.1}$$

where n is the maximum number of inference steps and f_t represents a fact retrieved at time-step t from the fact bank. We implement the model recursively by updating the hypothesis h at each time-step t, concatenating it with the partial explanation constructed at time step $t-1$:

$$h_t = g(h, f_1, \ldots f_{t-1}) \tag{5.2}$$

where $g(\cdot)$ represents the string concatenation function. The probability $P(f_t|h_t)$ is then approximated via an explanatory scoring function $es(\cdot)$ that jointly models *relevance* and *explanatory power* as:

$$es(f_t, h_t) = \lambda \cdot r(f_t, h_t) + (1-\lambda) \cdot pw(f_t, h) \tag{5.3}$$

where $r(\cdot)$ represents the relevance of f_t at time step t, while $pw(\cdot)$ represents the explanatory power of f_t.

As shown in Chapter 2 and 3, scientific explanations are composed of abstract sentences describing underlying explanatory laws and regularities that are frequently reused to explain a large set of hypotheses. To leverage this feature during inference,

we measure the explanatory power $pw(\cdot)$ of a fact as the extent to which it explains similar hypotheses in the explanations corpus.

The relevance $r(\cdot)$ is computed through a hybrid model that combines a sparse $s(\cdot)$ and a dense $d(\cdot)$ sentence encoder:

$$r(f_t, h_t) = sim(s(f_t), s(h_t)) + sim(d(f_t), d(h_t)) \tag{5.4}$$

with $sim(\cdot)$ representing the cosine similarity between two vectors. In our experiments, we adopt BM25 [129] as a sparse encoder, while Sentence-BERT [126] is adopted to train the dense encoder $d(\cdot)$.

5.3.1 Explanatory Power

Following Chapter 3, given a test hypothesis h, a sentence encoder $s(\cdot)$, and a corpus of scientific explanations, the explanatory power of a generic fact f_i can be estimated by analysing explanations for similar hypotheses in the corpus:

$$pw(f_i, h) = \sum_{h_k \in kNN(h)}^{K} sim(s(h), s(h_k)) \cdot \mathbb{1}(f_i, h_k) \tag{5.5}$$

$$\mathbb{1}(f_i, h_k) = \begin{cases} 1 & \text{if } f_i \in E_k \\ 0 & \text{if } f_i \notin E_k \end{cases} \tag{5.6}$$

where $kNN(h) = \{h_1, \ldots, h_K\}$ represents a list of hypotheses retrieved according to the similarity $sim(\cdot)$ between the embeddings $s(h)$ and $s(h_k)$, and $\mathbb{1}(\cdot)$ is the indicator function verifying whether f_i is part of the explanation E_k for the hypothesis h_k. Specifically, the more a fact f_i is reused for explaining hypotheses that are similar to h in the corpus, the higher its explanatory power.

In this work, we hypothesise that this model can be integrated within a hybrid framework based on dense and sparse encoders, improving inference performance while preserving scalability. In our experiments, we adopt BM25 similarity between hypotheses to compute the explanatory power efficiently.

5.3.2 Dense Bi-encoder

To train a dense encoder $d(\cdot)$, we fine-tune a Sentence-BERT model using a bi-encoder architecture [126]. The bi-encoder adopts a siamese network to learn a joint embedding space for hypotheses and facts in the fact bank. Following Sentence-BERT, we obtain

fixed sized sentence embeddings by adding a mean-pooling operation to the output vectors of BERT [39]. We employ a unique BERT model with shared parameters to learn a sentence encoder $d(\cdot)$ for both facts and hypotheses.

At the cost of sacrificing the performance gain resulting from self-attention, the bi-encoder allows for efficient multi-hop inference through Maximum Inner Product Search (MIPS). To enable scalability, we construct an index of dense embeddings for the whole fact bank. To this end, we adopt the approximated inner product search index (IndexIVFFlat) in FAISS [76].

5.3.3 Training

The bi-encoder is fine-tuned on inference chains extracted from annotated explanations in the WorldTree corpus [71]. Since the facts in the annotated explanations are not ordered, to train the model autoregressively, we first transform the explanations into sequences of facts sorting them in decreasing order of BM25 similarity with the hypothesis. We adopt BM25 since the facts that share less terms with the hypothesis tend to require more iterations and inference steps to be retrieved. Subsequently, given a training hypothesis h and an explanation sequence $E_{seq} = f_1, \ldots, f_n$, we derive n positive example tuples (h_t, f_t), one for each fact $f_t \in E_{seq}$, using $h_t = g(h, f_1, \ldots, f_{t-1})$ as hypothesis.

To make the model robust to distracting information, we construct a set of negative examples for each tuple (h_t, f_t) retrieving the top most similar facts to f_t that are not part of the explanation. We found that the best results are obtained using 5 negative examples for each positive tuple. We use the constructed training set and the siamese network to fine-tune the encoder via contrastive loss [52], which has been demonstrated to be effective for learning robust dense representations.

5.3.4 Multi-hop Inference

At each time-step t during inference time, we encode the concatenation of hypothesis and partial explanation h_t using the dense (Sentence-BERT) and sparse (BM25) encoders separately. Subsequently, we adopt the vectors representing h_t to compute the relevance score $r(\cdot)$ of the sentences in the fact bank (Equation 5.4). In parallel, the sparse representation (BM25) of the hypothesis h is adopted to retrieve the explanations for the top K similar hypotheses in the explanation corpus and compute the explanatory power $pw(\cdot)$ of each fact (Equation 5.5). Finally, relevance and explanatory power are

combined to compute the explanatory scores $es(\cdot)$ (Equation 5.3) and select the top candidate fact f_t from the fact bank to expand the explanation at time-step t. After t_{max} steps, we rank the remaining facts considering the explanation constructed a time-step t_{max}.

5.4 Empirical Evaluation

We perform an extensive evaluation on the WorldTree corpus adopting the dataset released for the shared task on multi-hop Explanation Regeneration[2] [71], where a diverse set of sparse and dense models have been evaluated. WorldTree is a subset of the ARC corpus [22] that consists of multiple-choice science questions annotated with natural language explanations supporting the correct answers. The WorldTree corpus provides a held-out test-set consisting of 1,240 science questions with masked explanations where we run the main experiment and comparison with published approaches.

To run our experiments, we first transform each question and correct answer pair into a hypothesis following the methodology described in [37]. We adopt explanations and hypotheses in the training-set ($\approx 1,000$) for training the dense encoder and computing the explanatory power for unseen hypotheses at inference time. We adopt `bert-base-uncased` [39] as a dense encoder to perform a fair comparison with existing cross-encoders employing the same model. The best results on Explanation Regeneration are obtained when running SCAR for 4 inference steps (additional details in Ablation Studies). In line with the shared task, the performance of the system is evaluated through the Mean Average Precision (MAP) of the produced ranking of facts with respect to the gold explanations in WorldTree. Implementation and pre-trained models adopted for the experiments are available online[3].

5.4.1 Explanation Regeneration

Table 5.1 reports the results achieved by our best model on the Explanation Regeneration task together with a comparison with previously published approaches. Specifically, we compare our hybrid framework based on bi-encoders with a variety of sparse and dense retrieval models.

Overall, we found that SCAR significantly outperforms all the considered sparse models (+5.39 MAP compared to the unification-based mechanism described in Chapter

[2] `https://github.com/umanlp/tg2019task`

[3] `https://github.com/ai-systems/hybrid_autoregressive_inference`

Model	Approach Description	MAP
Cross-encoders		
Cartuyvels et al. [19]	Autoregressive BERT	**57.07**
Das et al. [34]	BERT path-ranking + single fact ensemble	56.25
Das et al. [34]	BERT single fact	55.74
Das et al. [34]	BERT path-ranking	53.13
Chia et al. [21]	BERT re-ranking with gold IR scores	49.45
Banerjee [6]	BERT iterative re-ranking	41.30
Sparse Models		
Valentino et al. [161]	Unification-based Inference	**50.83**
Chia et al. [21]	Iterative BM25	45.76
Robertson et al. [129]	BM25 Relevance Score	43.01
Ramos et al. [124]	TF-IDF Relevance Score	39.42
Hybrid Models		
SCAR	Scalable Autoregressive Inference	**56.22**

Table 5.1: Results on the test-set and comparison with previous approaches. SCAR significantly outperforms all the sparse models and obtains comparable results with state-of-the-art cross-encoders.

3 [161]), obtaining, at the same time, comparable results with the state-of-the-art cross-encoder (-0.85 MAP compared to [19]). The following paragraphs provide a detailed comparison with previous work.

Dense Models. As illustrated in Table 5.1, all the considered dense models employ BERT [39] as a cross-encoder architecture. The state-of-the-art model proposed by [19] adopts an autoregressive formulation similar to SCAR. However, the use of cross-encoders makes the model computationally expensive and intrinsically not scalable. Due to the complexity of cross-encoders, in fact, the model can only be applied for re-ranking a small set of candidate facts at each iteration, which are retrieved using a pre-filtering step based on TF-IDF. In contrast, we found that the use of a hybrid model allows achieving comparable performance without cross-attention and pre-filtering step (-0.85 MAP), making SCAR approximately 50 times faster (see Section 5.4.2). The second-best dense approach employs an ensemble of two BERT models [34]. A first BERT model is trained to predict the relevance of each fact individually given a certain hypothesis. A second BERT model is adopted to re-rank a set of two-hops inference chains constructed via TF-IDF. The use of two BERT models in parallel, however,

Model	MAP ↑	Time (s/q) ↓
Autoregressive BERT	**57.07**	9.6
BERT single fact	55.74	18.4
BERT path-ranking	53.13	31.8
SCAR	56.22 (98.5%)	**0.19** (×50.5)

Table 5.2: Detailed comparison with BERT cross-encoders on the test-set in terms of Mean Average Precision (MAP – already introduced in Chapter 3) and inference time (seconds per question).

makes the approach computationally exhaustive. We observe that SCAR can achieve similar performance with the use of a single BERT bi-encoder, outperforming each individual sub-component in the ensemble with a drastic improvement in efficiency (SCAR is 96.8 times and 167.4 times faster, respectively, see Section 5.4.2). The remaining dense models [21, 6] adopt BERT-based cross-encoders to re-rank the list of candidate facts retrieved using sparse Information Retrieval (IR) techniques. As illustrated in Table 5.1, SCAR outperforms these approaches by a large margin (+6.77 and +14.92 MAP).

Sparse Models. We compare SCAR with sparse models presented on the Explanation Regeneration task. We observe that SCAR significantly outperforms the Unification-based Reconstruction model proposed in Chapter 3 [161] (+5.39 MAP), which employs a model of explanatory power in combination with BM25, but without dense representation and autoregressive inference. These results confirm the contribution of the hybrid model together with the importance of modelling Explanation Regeneration in a iterative fashion. In addition, we compare SCAR with the model proposed by [21] which adopts BM25 vectors to retrieve facts iteratively. We found that SCAR can improve the performance of this model by 10.46 MAP points. Finally, we measure the performance of standalone sparse baselines for a sanity check, showing that SCAR can significantly outperform BM25 and TFIDF (+13.21 and +16.8 MAP respectively), while preserving a similar level of scalability (see Sec. 5.4.6).

5.4.2 Inference Time

We performed additional experiments to evaluate the efficiency of SCAR and contrast it with state-of-the-art cross-encoders. To this end, we run SCAR on 1 16GB Nvidia Tesla

Model	t_{max}	**MAP**	**Time (s/q)**
Bi-encoder	1	41.98	0.04
	2	**42.17**	0.08
	3	39.97	0.12
	4	38.34	0.16
	5	37.24	0.19
	6	36.64	0.24
BM25	1	45.99	0.02
	2	47.77	0.04
	3	**48.35**	0.05
	4	48.06	0.07
	5	47.97	0.09
	6	47.66	0.11
Bi-encoder + BM25	1	51.53	0.05
	2	54.52	0.08
	3	55.65	0.14
	4	56.07	0.18
	5	**56.24**	0.22
	6	55.87	0.27
SCAR	1	57.10	0.06
	2	59.20	0.10
	3	59.73	0.15
	4	**60.28**	0.19
	5	59.79	0.24
	6	59.36	0.29

Table 5.3: Ablation study on the dev-set, where t_{max} represents the maximum number of iterations adopted to regenerate the explanations, and (s/q) is the inference time.

P100 GPU and compare the inference time with that of dense models executed on the same infrastructure [19]. Table 5.2 reports MAP and execution time in terms of seconds per question. As evident from the table, we found that SCAR is 50.5 times faster than the state-of-the-art cross-encoder [19], while achieving 98.5% of its performance. Moreover, when compared to the individual BERT models proposed by [34], SCAR is able to achieve better MAP score (+0.48 and +3.09), increasing even more the gap in terms of inference time (96.8 and 167.4 times faster).

5.4.3 Ablation Studies

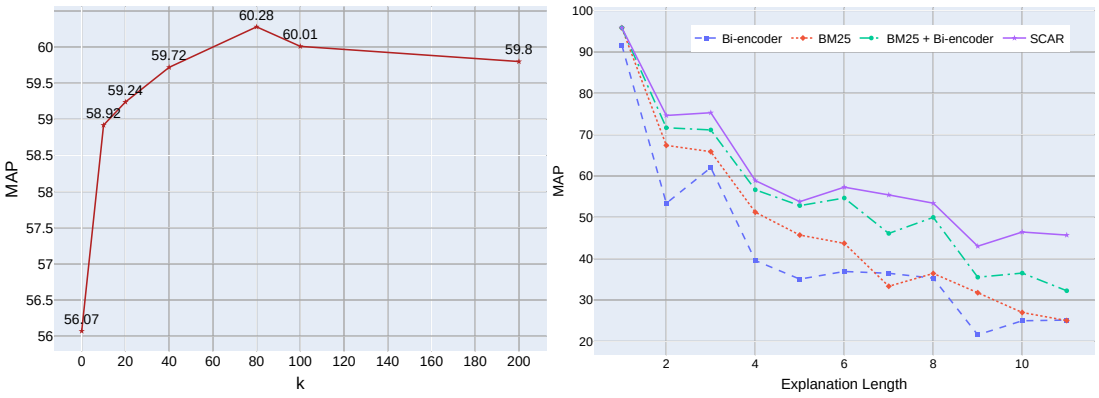
In order to understand how the different components of SCAR complement each other, we carried out distinct ablation studies. The studies are performed on the dev-set since the explanations on the test-set are masked.

Table 5.3 presents the results on Explanation Regeneration for different ablations of SCAR adopting an increasing number of iterations t_{max} for the inference. The results show how the performance improves as we combine sparse and dense models, with a decisive contribution coming from each individual sub-component. Specifically, considering the best results obtained in each case, we observe that SCAR achieves an improvement of 18.11 MAP over the dense component (Bi-encoder) and 11.93 MAP when compared to the sparse model (BM25). Moreover, the ablation demonstrates the fundamental role of the explanatory power model in achieving the final performance, which leads to an improvement of 4.04 MAP over the Bi-encoder + BM25 model (Equation 5.3).

Overall, we notice that performing inference iteratively is beneficial to the performance across the different components. We observe that the improvement is more prominent when comparing $t_{max} = 1$ (only using the hypothesis) with $t_{max} = 2$ (using hypothesis and first fact), highlighting the central significance of the first retrieved fact to support the complete regeneration process. Except for the Bi-encoder, the experiments demonstrate a slight improvement when adding more iterations to the process, obtaining the best results for SCAR using a total of 4 inference steps.

We notice that the best performing component in terms of inference time is BM25. The integration with the dense model, in fact, slightly increases the inference time, yet leading to a decisive improvement in terms of MAP score. Even with the overhead caused by the Bi-encoder, however, SCAR can still perform inference in less than half a second per question, a feature that demonstrates the scalability of the approach with respect to the number of iterations.

Finally, we evaluate the impact of the explanatory power model by considering a larger set of training hypotheses for its implementation (Figure 5.3a). To this end, we compare the performance across different configurations with increasing values of K in Equation 5.5. The results demonstrate the positive impact of the explanatory power model on the inference, with a rapid increase of MAP peaking at $K = 80$. After reaching this value, we observe that considering additional hypotheses in the corpus has little impact on the model's performance.



(a) (b)

Figure 5.3: (a) Impact of increasing the number of similar hypotheses K to estimate the explanatory power (Equation 5.5). (b) Performance considering hypotheses with gold explanations including an increasing number of facts.

5.4.4 Semantic Drift

Recent work have shown that the regeneration of scientific explanations is particularly challenging for multi-hop inference models as it can lead to a phenomenon known as *semantic drift* – i.e., the composition of spurious inference chains caused by the tendency of drifting away from the original context in the hypothesis [82, 180, 71, 151]. In general, the larger the size of the explanation, the higher the probability of semantic drift. Therefore, it is particularly important to evaluate and compare the robustness of multi-hop inference models on hypotheses requiring long explanations. To this end, we present a study of semantic drift, comparing the performance of different ablations of SCAR on hypotheses with a varying number of facts in the gold explanations.

The results of the study are reported in Figure 5.3b. Overall, we observe a degradation in performance for all the considered models that becomes more prominent as the explanations increase in size. Such a degradation is likely due to semantic drift. However, the results suggest that SCAR exhibits more stable performance on long explanations ($\geq$ 6 facts) when compared to its individual sub-components. In particular, the plotted results in Figure 5.3b clearly show that, while all the models start with comparable MAP scores on explanations containing a single fact, the gap in performance gradually increases with the size of the explanations, with SCAR obtaining an improvement of 13.46 MAP over BM25 + Bi-encoder on explanations containing more than 10 facts. These results confirm the hypotheses that implicit and explicit patterns possess complementary features for Explanation Regeneration and that the proposed

Model	t = 1	t = 2	t = 3	t = 4
Random	25.00	25.00	25.00	25.00
BM25	**48.23**	39.82	35.84	33.18
Bi-encoder	**54.42**	52.21	50.88	50.00
Bi-encoder + BM25	**59.29**	52.21	47.79	44.69
SCAR	60.62	60.62	**61.06**	57.96

Table 5.4: Accuracy in question answering using the models as explanation-based inference solvers without additional training.

hybridisation has a decisive impact on improving multi-hop inference for scientific hypotheses in the most challenging setting.

5.4.5 Multi-hop Question Answering

Since the construction of spurious inference chains can lead to wrong answer prediction, semantic drift often influences the downstream capabilities of answering the question. Therefore, we additionally evaluate the performance of SCAR on the multiple-choice question answering task (WorldTree dev-set), employing the model as an explanation-based solver without additional training. Specifically, given a multiple-choice science question, we employ SCAR to construct an explanation for each candidate answer, and derive the relative candidate answer score by summing up the explanatory score of each fact in the explanation (Equation 5.3). Subsequently, we consider the answer with the highest-scoring explanation as the correct one.

Table 5.4 shows the results achieved adopting different iterations t for the inference. Similarly to the results on Explanation Regeneration, this experiment confirms the interplay between dense and sparse models in improving the performance and robustness on downstream question answering. Specifically, we observe that, while the performance of different ablations decreases rapidly with an increasing number of inference steps, the performance of SCAR are more stable, reaching a peak at $t = 3$. This confirms the robustness of SCAR in multi-hop inference together with its resilience to semantic drift.

5.4.6 Scalability

We measure the scalability of SCAR on fact banks containing milions of sentences. To perform this analysis, we gradually expand the set of facts in the WorldTree corpus

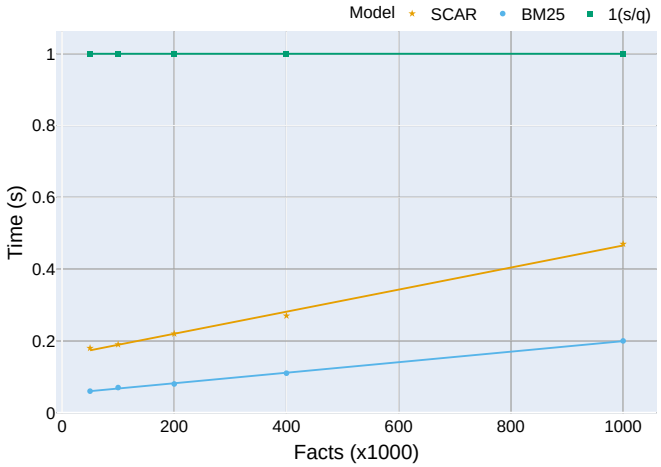


Figure 5.4: Scalability of SCAR to corpora containing a million facts compared to that of standalone BM25.

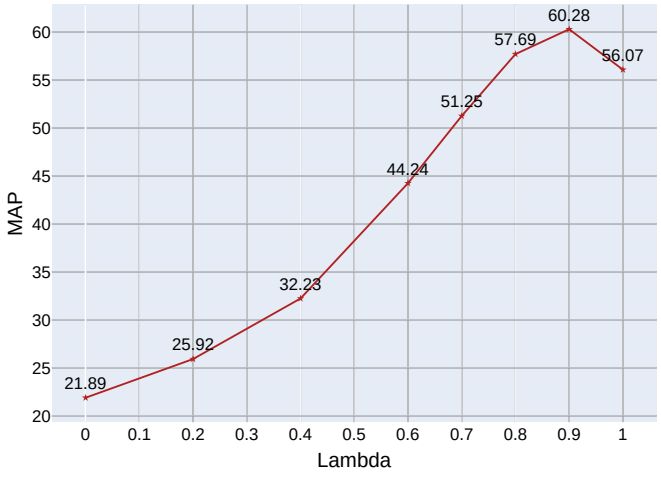


Figure 5.5: Tuning of λ for the explanatory scoring function $es(\cdot)$ (Equation 5.3).

by randomly extracting sentences from GenericsKB[4] [11], a curated fact bank of commonsense and scientific knowledge. To evaluate scalability, we compare the inference time of SCAR with that of standalone BM25, which is widely adopted for Information Retrieval at scale [129].

The results of this experiment, reported in Figure 5.4, demonstrate that SCAR scales similarly to BM25. Even considering the overhead caused by the Bi-encoder model, in fact, SCAR is still able to perform inference in less than 1 second per question on corpora containing 1 million facts, demonstrating its suitability for scenarios requiring inference on large knowledge sources.

[4] https://allenai.org/data/genericskb

5.4.7 Sensitivity Analysis

The hyperparameter λ in Equation 5.3 has been tuned to maximise the MAP for Explanation Regeneration on the WorldTree dev-set. Specifically, we found that the best results are obtained for $\lambda = 0.89$.

Figure 5.5 shows the MAP score obtained with different values of λ, with 0 representing the extreme case in which only the explanatory power is active and 1 the case in which only the relevance score is active. As shown in the graph, the explanatory power alone does not allow achieving high performance on the task, demonstrating that explanations for unseen hypotheses cannot be simply regenerated considering similar hypotheses in the training set and that the relevance model is necessary for generalisation on unseen problems.

5.5 Related Work

Multi-hop inference is the task of combining multiple pieces of evidence to solve a particular reasoning problem. This task is often used to evaluate explanation-based inference since the constructed chains of reasoning can be interpreted as an explanation for the final predictions [173, 151]. Given the importance of multi-hop reasoning for explainability, there is a recent focus on resources providing annotated explanations to support the inference [180, 74, 88, 44, 83, 187, 111, 72, 171]. Most of the existing datasets for multi-hop inference, however, contain explanations composed of up to two sentences or paragraphs, limiting the possibility to assess the robustness of the systems on long reasoning chains. Consequently, most of the existing multi-hop models are evaluated on two-hops inference tasks [103, 156, 186, 43, 4, 7, 185, 155].

Explanation Regeneration. Explanation Regeneration on science questions is designed to evaluate the construction of long explanatory chains in a setting where the structure of the inference cannot be derived from a direct decomposition of the questions [180, 71, 72]. To deal with the difficulty of the task, state-of-the-art models leverage the attention mechanism in Transformers [163], learning to compose relevant explanatory chains via sequence classification models [19, 34, 21, 6]. The autoregressive formulation proposed in this chapter is similar to the one introduced by [19], which, however, perform iterative inference though a cross-encoder architecture based on BERT [39]. Differently from this work, we present a hybrid architecture based on bi-encoders

[126] with the aim of optimising both accuracy and inference time in Explanation Regeneration.

Multi-hop Inference with Dense Retrieval Our framework is related to recent work on dense retrieval for knowledge-intensive NLP tasks, which focuses on the design of scalable architectures with Maximum Inner Product Search (MIPS) based on Transformers [181, 189, 104, 77, 100, 40]. Our multi-hop dense encoder is similar to [104] and [181] which adopt bi-encoders for multi-step retrieval on open-ended commonsense reasoning and open-domain question answering. However, to the best of our knowledge, we are the first to integrate dense bi-encoders in a hybrid architecture for complex explanation-based inference in the scientific domain.

5.6 Conclusion

This work presented SCAR, a hybrid autoregressive architecture for scalable Explanation Regeneration. An extensive evaluation demonstrated that SCAR achieves performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster and intrinsically scalable, confirming the impact of the hybridisation on semantic drift and question answering. This work demonstrated the effectiveness of hybrid architectures for explanation-based inference at scale, opening the way for future research at the intersection of latent and explicit models. As a future work, we plan to investigate the integration of relevance and explanatory power in an end-to-end differentiable architecture, and explore the applicability of the hybrid framework on additional natural language and scientific reasoning tasks, with a focus on real-world scientific inference problems.

5.7 Scoping and Limitations

The autoregressive inference model assumes that the relevance of a fact at each time step t depends on the whole explanation sequence constructed at time $t-1$. However, while it is reasonable to construct explanations in an iterative fashion, the notion of relevance might depend only on a subpart of the partially constructed explanation. In that sense, the relevance dependencies between sentences in the explanation might actually induce a sparse graphical structure. While cross-encoders can model these dependencies through the self-attention mechanism in Transformers, the same mechanism cannot be

leveraged using bi-encoders. Therefore, additional work is required for modelling this feature of explanatory relevance, including the investigation of different forms of representations (e.g. graph-based), the adoption of sparse concatenation mechanisms during autoregressive inference, and the exploration of hybrid scalable solutions preserving self-attention such as poly-encoders [63].

Chapter 6

Explanation Gold Standards

An emerging line of research in Explanation-based NLI is the creation of datasets enriched with human-annotated explanations and rationales, used to build and evaluate models with step-wise inference and explanation generation capabilities. While human-annotated explanations are used as ground-truth for the inference, there is a lack of systematic assessment of their consistency and rigour. In an attempt to provide a critical quality assessment of Explanation Gold Standards (XGSs) for NLI, this chapter investigates ***RQ5:*** *"Do natural language explanations in existing gold standards represent valid and complete logical arguments?"* proposing a systematic annotation methodology named *Explanation Entailment Verification* (*EEV*).

The application of *EEV* on three mainstream datasets reveals the conclusion that a majority of the explanations, while appearing coherent on the surface, represent logically invalid arguments, ranging from being incomplete to containing identifiable logical errors. This conclusion confirms that the inferential properties of explanations are still poorly formalised and understood, and that additional work on this line of research is necessary to improve the way Explanation Gold Standards are constructed[1].

6.1 Introduction

Explanation Gold Standards (XGSs) are emerging as a fundamental enabling tool for step-wise and Explaination-based Natural Language Inference (NLI). Resources such as WorldTree [179, 72], QASC [87], among others [173, 151, 10, 15] provide a corpus of linguistic evidence on how humans construct explanations that are perceived as

[1]This chapter follows the publication "Do Natural Language Explanations Represent Valid Logical Arguments? Verifying Entailment in Explainable NLI Gold Standards"[158]

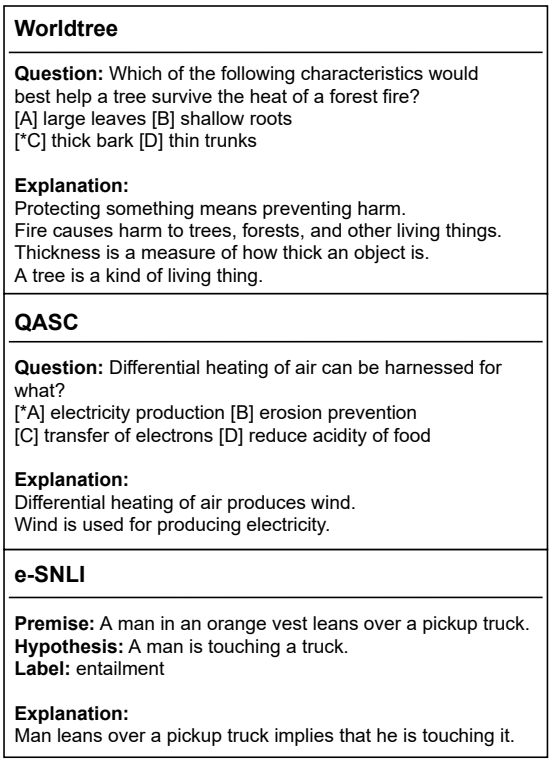


Figure 6.1: Does the answer logically follow from the explanation? While step-wise explanations are used as ground-truth for the inference, there is a lack of assessment of their consistency and rigour. We propose *EEV*, a methodology to quantify the logical validity of human-annotated explanations.

plausible, coherent and complete.

Designed for tasks such as Textual Entailment (TE) and Question Answering (QA), these reference datasets are used to build and evaluate models with step-wise inference and explanation generation capabilities [161, 18, 93, 123]. While these explanations are used as ground-truth for the inference, there is a lack of systematic assessment of their consistency and rigour, introducing inconsistency biases within the models.

This chapter aims to provide a critical quality assessment of Explanation Gold Standards for NLI in terms of their logical inference properties. By systematically translating natural language explanations into corresponding logical forms, we induce a set of recurring logical violations which can then be used as testing conditions for quantifying quality and logical consistency in the annotated explanations. More fundamentally, the chapter reveals the conclusion that a majority of the explanations present in existing gold standards contain one or more major logical fallacies, while appearing to be coherent on the surface.

The main contributions of this chapter can be summarised as:

1. Proposal of a systematic methodology, named *Explanation Entailment Verification* (*EEV*), for analysing the logical consistency of NLI explanation gold-standards.

2. Validation of the quality assessment methodology for three contemporary and mainstream reference XGSs.

3. The conclusion that most of the annotated human-explanations in the analysed samples represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors.

6.2 Explanation Gold Standards

Given a generic classification task T, an Explanation Gold Standard (XGS) is a collection of distinct instances of T, $XGS(T) = \{I_1, I_2, \ldots, I_n\}$, where each element of the set, $I_i = \{X_i, s_i, E_i\}$, includes a problem formulation X_i, the expected solution s_i for X_i, and a human-annotated explanation E_i.

In general, the nature of the elements in a XGS can vary greatly according to the task T under consideration. In this work, we restrict our investigation to Natural Language Inference (NLI) tasks, such as Textual Entailment and Question Answering, where problem formulation, expected solution, and explanations are entirely expressed in natural language.

For this class of problems, the explanation is typically a composition of sentences, whose role is to describe the reasoning required to arrive at the final solution. As shown in the examples depicted in Figure 6.1, the explanations are constructed by human annotators transcribing the commonsense and world knowledge necessary for the correct answer to hold. Given the nature of XGSs for NLI, we hypothesise that a human-annotated explanation represents a valid set of premises from which the expected solution logically follows.

Specifically, to investigate ***RQ5:*** *"Do natural language explanations in existing gold standards represent valid and complete logical arguments?"*, we present a methodology that is guided by the following research hypothesis:

- **RH5.1**: human-annotated explanations represent valid and complete arguments from which the solution for a given NLI problem logically follows.

In order to validate or reject this hypothesis, we design a methodology aimed at evaluating XGSs in terms of logical entailment, quantifying the extent to which human-annotated explanations actually entail the final answer.

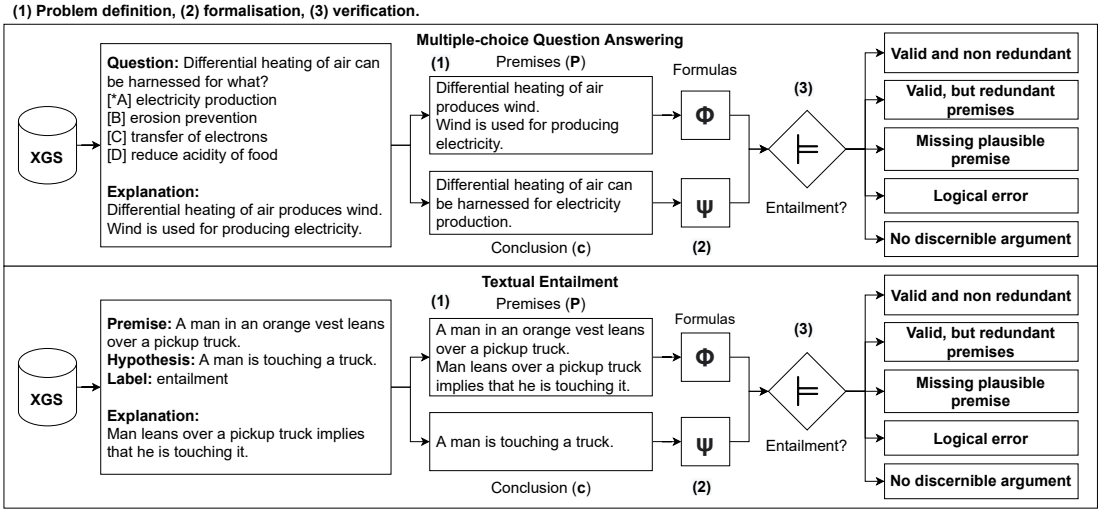


Figure 6.2: Overview of the Explanation Entailment Verification (*EEV*) applied to different NLI problems. *EEV* takes the form of a multi-label classification problem where, for a given NLI problem, a human annotator has to qualify the validity of the inference process described in the explanation through a pre-defined set of classes.

6.3 Explanation Entailment Verification

We present an annotation framework, named Explanation Entailment Verification (*EEV*), that takes the form of a multi-label classification problem defined on a XGS. Specifically, the goal of *EEV* is to label each element in a XGS, $I_i = \{X_i, s_i, E_i\}$, using one of a predefined set of classes qualifying the validity of the inference process described in the explanation E_i.

Figure 6.2 shows a schematic representation of the annotation pipeline. One of the challenges involved in the design of a standardised methodology for *EEV* is the formalisation of an annotation task that is applicable to NLI problems with different shapes, such as Textual Entailment (TE) and Multiple-choice Question Answering (MCQA). To minimise the ambiguity in the annotation and make it independent of the specific NLI task, we define a methodology composed of three major steps: (1) *problem definition*; (2) *formalisation*; and (3) *verification*.

In the problem definition step, each example I_i in the XGS is translated into an entailment form ($P \models c$), identifying a set of sentences P representing the premises for the entailment, and a single sentence c representing its conclusion. As illustrated in Figure 6.2, this step defines an entailment problem with a single surface form that allows abstracting from the NLI task under investigation.

In the formalisation step, the sentences in P and c are translated into a logical

form ($\Phi \models \psi$) (additional details on the formalism are provided in section 6.3.3). This step aims to minimise the ambiguity in the interpretation of the meaning of the sentences, supporting the annotators in the identification of logical errors and gaps in the explanations, and maximise the inter-annotator agreement in the downstream verification task.

The final step corresponds to the actual multi-label classification problem. Specifically, the annotators are asked to verify whether the formalised set of premises Φ entails the conclusion ψ ($\Phi \models \psi$) and to classify the explanation in the corresponding example $I_i = \{X_i, s_i, E_i\}$ selecting one of the following classes: (1) *Valid and non redundant*; (2) *Valid, but redundant premises*; (3) *Missing plausible premise*; (4) *Logical error*; (5) *No discernible argument*. The classes are mutually exclusive: each example can be assigned to one and only one label.

After *EEV* is performed for each instance in the dataset, the frequencies of the classification labels can be adopted to estimate and evaluate the overall entailment properties of the explanations in the XGS under consideration.

6.3.1 Problem definition

The problem definition step consists in the identification of the sentences in $I_i = \{X_i, s_i, E_i\}$ that will compose the set of premises P and the conclusion c for the entailment problem $P \models c$.

Here, we describe the procedure adopted for translating a specific NLI task into the entailment problem of interest given its original surface form. In particular, we employ two different translation procedures for Textual Entailment (TE) and Multiple-choice Question Answering (MCQA) problems.

Textual Entailment (TE). For a TE task, the problem formulation X_i is generally composed of two sentences, p and h, representing a premise and a hypothesis (see e-SNLI in figure 6.1). Each example in a TE task can be classified using one of the following labels: *entailment*, *neutral*, and *contradiction* [13]. In this work, we focus on examples where the expected solution s_i is *entailment*, implying that the hypothesis h is a consequence of the premise p. Therefore, to define the entailment verification problem, we simply include the premise p in P and consider the hypothesis h as a the conclusion c. For this class of problems, the explanation E_i describes additional factual knowledge necessary for the entailment $p \models h$ to hold [15]. Specifically, the sentences

in E_i can be interpreted as a further set of premises for the entailment verification problem and are included in P.

Multiple-choice Question Answering (MCQA). In the case of MCQA, X_i is typically composed of a question $Q_i = \{c_1, \ldots, c_n, q\}$, and a set of mutually exclusive candidate answers $A_i = \{a_1, \ldots, a_m\}$ (see QASC and WorldTree in figure 6.1). In this case, the expected label s_i corresponds to one of the candidate answers in A_i [72, 87]. Q_i can include a set of introductory sentences $c_1, \ldots, c_n$ acting as a context for the question q. We consider each sentence c_i in the context as a premise for q and include it in P. Similarly to TE, we interpret the explanation E_i for a MCQA example as a set of premises that entails the correct answer s_i. Therefore, the sentences in E_i are included in P. The question q takes the form of an elliptical assertion, and the candidate answers are possible substitutions for the ellipsis. Therefore, to derive the conclusion c, we adopt the correct answer s_i as a substitution for the ellipsis in q. Details on the formalisation adopted for MCQA problems are described in section 6.3.3.

6.3.2 Verification

In the verification step, the annotators adopt the formalised set of premises Φ and conclusion ψ to classify the entailment problem in one of the following categories:

1. **Valid and non-redundant:** The argument is formally valid, and all premises are required for the derivation.

2. **Valid, but redundant premises:** The argument is formally valid, but some premises are not required for the derivation. This includes the cases where more than one premise is present, and the conclusion simply repeats one of the premises.

3. **Missing plausible premise:** The argument is formally invalid, but would become valid on addition of a reasonable premise, such as, for example, *"If x affects y, then a change to x affects y"*, or *"If x is the same height as y and y is not as tall as z then x is not as tall as z"*.

4. **Logical error:** The argument is formally invalid, apparently as a result of confusing *"and"* and *"or"* or *"some"* and *"all"*, or of illicitly changing the direction of an implication.

5. **No discernible argument:** The argument is invalid, no obvious rescue exists in the form of a missing premise, and no simple logical error can be identified.

6.3.3 Formalisation

In this section, we describe an example of formalisation for a MCQA problem. A typical multiple-choice problem is a triple consisting of a *question* Q together with a set of *candidate answers* $A_1, \ldots, A_m$. It is understood that Q takes the form of a elliptical assertion, and the candidate answers are possible substitutions for the ellipsis. The corpora investigated feature a list of multiple-choice textual entailment problems together, in each case, with a specification of a correct answer and an *explanation* in the form of a set of assertions Φ providing a justification for the answer. For example, the following problem together with its resolution is taken from the WorldTree corpus [72].

Question: A group of students are studying bean plants. All of the following traits are affected by changes in the environment except ...

Candidate answers: [A] leaf color. [B] seed type. [C] bean production. [D] plant height.

Correct answer: B

Explanation: (i) The type of seed of a plant is an inherited characteristic; (ii) Inherited characteristics are the opposite of learned characteristics; acquired characteristics; (iii) An organism's environment affects that organism's acquired characteristics; (iv) A plant is a kind of organism; (v) A bean plant is a kind of plant; (vi) Trait is synonymous with characteristic.

In formalising such problems, we represent the question as a sentence of first-order logic featuring a schematic formula variable P (corresponding to the ellipsis), and the candidate answers as first-order formulas. In the above example, we assume that the essential force of the question to find a characteristic of plants *not* affected by those

plants' environments. That is, we are asked for a P making the schematic formula

$$\forall xyzwe(\text{bnPlnt}(x) \wedge \text{env}(y,x) \wedge \text{changeIn}(z,y) \wedge \text{trait}(w,x) \wedge \text{affct}(e) \wedge \text{agnt}(e,z) \wedge P \rightarrow \neg\text{ptnt}(e,w)). \quad (6.1)$$

into a true statement. We formalise the correct answer (B) by the atomic formula $\text{sdTp}(w,x)$ "w is the seed type of x", with the other candidate answers formalised similarly. In choosing predicates for formalisation, we typically render common noun-phrases using predicates, taking these to be relational if the context demands (e.g. "environment/seed type *of* a plant x"). In addition, we typically render verbs as predicates whose arguments range over eventualities (events, processes, etc.), related to their participants via a standard list of binary "semantic role" predicates (agent, patient, theme) etc. Thus, to say that "x affects y" is to report the existence of an eventuality e of type "affecting", such that x is the agent of e and y its patient. This approach, although somewhat strained in many general contexts, aids standardization and, more importantly, also makes it easier to deal with adverbial phrases. Of course, many choices in formalisation strategy inevitably remain.

The explanation Φ is formalised as a finite set of first-order formulas, following the same general rendering policies. In the case of the above example, sentences (i), (ii) and (iv)–(vi) in Φ might be formalised as:

$$\begin{aligned}
&\forall xy(\text{plnt}(x) \wedge \text{sdTp}(y,x) \rightarrow \text{char}(y,x) \wedge \text{inhtd}(y)) \\
&\forall xy(\text{char}(x,y) \wedge \text{inhtd}(x) \rightarrow \neg\text{acqrd}(x)) \\
&\forall x(\text{plnt}(x) \rightarrow \text{orgnsm}(x)) \\
&\forall x(\text{bnPlnt}(x) \rightarrow \text{plnt}(x)) \\
&\forall xy(\text{trait}(x,y) \leftrightarrow \text{char}(x,y)),
\end{aligned}$$

with the more complicated sentence (iii) formalised as

$$\forall xyw(\text{orgnsm}(x) \wedge \text{env}(y,x) \wedge \text{char}(w,x) \wedge \text{acqrd}(w) \rightarrow \exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w))) \quad (6.2)$$

Denoting by ψ the result of substituting $\text{sdTp}(w,x)$ for P in (6.1), we ask ourselves:

Does Φ entail ψ? A moment's thought shows that it does not. At the very least, statement (iii) in the explanation, whose formalisation is (6.2), must instead be read as asserting that an organism's environment affects *only* that organism's acquired characteristics, that is to say:

$$\begin{aligned}&\forall xyw(\text{orgnsm}(x) \wedge \text{env}(y,x) \wedge \text{char}(w,x) \wedge \\ &\exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w)) \rightarrow \\ &\qquad \text{acqrd}(w)).\end{aligned} \tag{6.3}$$

This is not unreasonable, of course. Generalizations in natural language are notoriously vague as to the direction of implication; let Φ' be the result of substituting (6.3) for (6.2) in Φ. Does Φ' entail ψ? Again, no. The problem this time is that, model-theoretically speaking, just because something is affected by a *change in* its environment, that does not mean to say it is affected by its environment. An assertion to the effect that it is would have to be postulated:

$$\begin{aligned}&\forall xyzw(\text{env}(y,x) \wedge \text{changeIn}(z,y) \wedge \\ &\quad \exists e(\text{affct}(e) \wedge \text{agnt}(e,z) \wedge \text{ptnt}(e,w)) \rightarrow \\ &\quad \exists e(\text{affct}(e) \wedge \text{agnt}(e,y) \wedge \text{ptnt}(e,w))).\end{aligned}$$

Let Φ'' be the result of augmenting Φ' in this way. Then Φ'' does indeed entail ψ. Therefore, we classify this example as a *missing plausible premise*.

6.4 Corpus Analysis

We employ *EEV* to analyse a set of contemporary XGSs designed for Textual Entailment and Multiple-choice Question Answering.

In the following sections, we describe the methodology adopted for extracting a representative sample from the selected XGSs, and for implementing the annotation pipeline efficiently. Finally, we present the results of the annotation, reporting the frequency of each entailment verification class and presenting a list of qualitative examples to provide additional insights on the logical properties of the analysed explanations.

Feature	Worldtree	QASC	e-SNLI
Task	MCQA	MCQA	TE
Multi-hop	yes	yes	no
Crowd-sourced	no	yes	yes
Explanation type	generated + composed	composed	generated
Avg. number of sentences	6	2	1

Table 6.1: Features of the datasets selected for the Explanation Entailment Verification (*EEV*).

6.4.1 Selected Datasets

We select three contemporary XGSs with different and complementary characteristics. In particular, we apply our methodology to two MCQA datasets (WorldTree [72], QASC[87]) and one TE benchmark (e-SNLI [15]).

The main features of the selected XGSs are reported in Table 6.1. *Multi-hop* indicates whether the problem requires step-wise reasoning, combining more than one sentence to compose the final explanation. *Crowd-sourced* indicates whether the resource is curated using standard crowd-sourcing platforms. *Explanation type* represents the methodology adopted to construct the explanations. *Generated* means that the sentences in the explanations are entirely created by human annotators. On the other hand, *composed* means that the sentences are retrieved from an external knowledge resource. Finally, the last row reports the *average number of sentences* composing the explanations.

6.4.2 Annotation Task

The bottleneck of the annotation framework lies in the formalisation phase, which is generally time consuming and requires trained experts in the field. In order to make the application of *EEV* efficient in practice, we extract a sub-set of $n = 100$ examples from each XGS (Worldtree, QASC, and e-SNLI). To maximise the representativeness of the explanations in the subset, given a fixed size n, we combine a set of sampling methodologies with effect size analysis. The details of the sampling methodology are described in section 6.4.3 while the results are presented in section 6.4.4. Code and data adopted for the experiments are available online [2].

The extracted examples are randomly assigned to 2 annotators with an overlap of

[2] https://github.com/ai-systems/explanation-entailment-verification/

20 instances to compute the inter-annotator agreement. All the annotators are active researchers in the field of Natural Language Processing and Computational Semantics. Table 6.2 reports the inter-annotator agreement achieved on each dataset separately. Overall, we observe an average of 72% accuracy in the multi-label classification task, computed considering the percentage of overlaps between the final entailment verification classes chosen by the annotators.

6.4.3 Sampling Methodology

To maximise the representativeness of the explanations for the subsequent annotation task, while analysing a fixed number n of examples for each dataset, we combine a set of sampling methodologies with effect size analysis. In this section, we describe the sampling techniques adopted for each dataset.

A stratified sampling methodology has been adopted for the WorldTree corpus [179, 72]. The stratified sampling consists in partitioning the dataset using a set of classes and performing random sampling from each class independently. This strategy guarantees that the same amount of examples is extracted from each class. The stratified technique requires the classes to be collectively exhaustive and mutually exclusive – i.e, each example has to belong to one and only one class. To apply stratified sampling on Worldtree, we consider the high-level topics introduced in [182], which are used to classify each question in the dataset according to one of the following categories: Life, Earth, Forces, Materials, Energy, Scientific Inference, Celestial Objects, Safety, Other. The same technique cannot be applied to e-SNLI [15] and QASC [87] since the examples in these datasets are not partitioned using any abstract set of classes. In this case, therefore, we use random sampling on the whole dataset to extract a fixed number n of examples.

Once a fixed number of examples n is extracted from each dataset, we consider the annotated explanation sentences of each example to verify whether the extracted set of explanations is representative of the whole dataset. To perform this analysis, we assume the predicates in the explanation sentences to be the expression of the type of knowledge of the whole explanation. Therefore, we consider the extracted sample of explanations representative if the distribution of predicates in the sample is correlated with the same distribution in the whole dataset. To this end, we compute the frequencies of the verbs appearing in the explanation sentences from the extracted sub-set and original dataset separately. Subsequently, we compare the frequencies in the sub-sample with the frequencies in the whole dataset computing a Pearson correlation coefficient.

Dataset	Agreement Accuracy
Worldtree	.70
QASC	.70
e-SNLI	.75

Table 6.2: Inter-annotator agreement computed in terms of accuracy in the multi-label classification task considering the first annotator as a gold standard.

Dataset	Correlation Coefficient
Worldtree	.964
QASC	.958
e-SNLI	.987

Table 6.3: Effect size analysis of the samples extracted from each XGS for the downstream EEV annotation.

In this case, a coefficient greater than .7 indicates a strong correlation between the types of explanations in the sample and the types of explanations in the original dataset. After running the sampling for t times independently, we select the subset of explanations for each dataset with the highest Pearson correlation coefficient. Table 6.3 reports the Pearson correlation for the subsets adopted in our analysis with fixed sample size $n = 100$.

6.4.4 Results

The quantitative analysis presented in this section aims to empirically assess the hypothesis that human-annotated explanations in XGSs constitute *valid and complete* logical arguments for the expected answers. We report the quantitative results of the explanation entailment verification in Table 6.4. Specifically, the table reports the percentage of the frequency of each verification class in the analysed samples. The column *AVG* reports the average for each class.

Overall, we observe that the results of the annotation task tend to reject our research hypothesis, with an average of only 20.42% of analysed explanations being classified as *valid and non redundant* arguments. When considering also *valid, but redundant* explanations (21.91%), the average percentage of valid arguments reaches a total of 42.33%. Therefore, we can conclude that the majority of the explanations represent formally invalid arguments (57.66%).

We observed that the majority of invalid arguments are classified as *missing plausible*

Entailment Verification Class	Worldtree	QASC	e-SNLI	AVG
Valid and non-redundant	12.24	17.65	**31.37**	20.42
Valid, but redundant premises	26.53	7.84	**31.37**	21.91
Missing plausible premise	**38.78**	21.57	17.65	**26.00**
Logical error	6.12	17.65	9.80	11.19
No discernible argument	16.33	**35.29**	9.80	20.47
Valid argument	38.77	25.49	**62.74**	42.33
Invalid argument	61.23	**74.51**	37.25	**57.66**

Table 6.4: Results of the application of *EEV* for each entailment verification category.

premise. This finding implies that a significant percentage of annotated explanations are incomplete arguments (26.00%), that can be made valid on addition of a reasonable premise. We attribute this result to the tendency of human explainers to take for granted part of the world knowledge required for the entailment to hold [165].

A lower but significant percentage of explanations contain identifiable logical errors (11.19%), which result from confusing the set of quantifiers and logical operators, or from illicitly changing the direction of an implication. Similarly, 20.47% of the explanations where labeled as *no discernible arguments*, where no obvious premise can be added to make the argument valid and no simple logical error can be detected. This result can be attributed partly to natural errors occurring in a gold standard creation process, partly to the effort required for human-annotators to identify logical fallacies in their explanations. In the remaining of this section, we analyse the results obtained on each XGS.

WorldTree. The analysed sample contains the highest percentage of incomplete arguments, with a total of 38.78% explanations classified as *missing plausible premise*. This result can be explained by the fact that the questions in WorldTree require complex forms of reasoning, facilitating the construction of arguments containing implicit world knowledge and missing premises. At the same time, the dataset contains the smallest percentage of logical errors (6.12%). We attribute this outcome to the fact that WorldTree is not crowd-sourced, implying that the quality of the annotated explanations is more easily controllable using internal verification methods.

QASC. This XGS contains the highest rate of invalid arguments (62.74%), with 35.29% of the explanations classified as *no discernible argument*. One of the factors

Problem Formulation	Explanation	XGS
Valid and non-redundant (20.42%)		
Premise: A smiling woman is playing the violin in front of a turquoise background. **Hypothesis:** A woman is playing an instrument.	A violin is an instrument.	e-SNLI
Valid, but redundant premises (21.91%)		
Premise: Four people are bandaging a head wound. **Hypothesis:** People are bandaging an injured head.	People are bandaging an injured head wound.	e-SNLI
Missing plausible premise (26.00%)		
Question: A group of students are studying bean plants. All of the following traits are affected by changes in the environment except [A] Leaf color [*B] Seed type [C] Bean production [D] Plant height	The type of seed of a plant is an inherited characteristic. Inherited characteristics are the opposite of learned characteristics; acquired characteristics. An organism's environment affects that organism's acquired characteristics. A plant is a kind of organism. Trait is synonymous with characteristic.	Worldtree
Logical error (11.19%)		
Question: What can use energy in order to make food for itself and others? [A] Organisms [B] Mollusks [C] Trees [D] Microbes [E] Seeds [F] Chlorophyll [*G] Plants [H] Animals	Producers use energy and inorganic molecules to make food. If the plant is green, it is a producer.	QASC
No discernible argument (20.47%)		
Question: What converts mechanical energy into kinetic energy when it moves? [*A] dogs [B] Bats [C] camels [D] Birds [E] Mammal [F] bears [G] hawks [H] Whales	When an animal moves, chemical energy is converted to mechanical energy. All dogs are animals.	QASC

Table 6.5: Examples of explanations classified with different entailment verification categories.

contributing to these results might be related to the length of the constructed explanations, which is limited to 2 facts extracted from a predefined corpus of sentences. The high rate of no discernible arguments and missing premises (35.29% and 21.57% respectively) suggests that the majority of the questions require additional world knowledge and more detailed explanations. This conclusion is also supported by the percentage of *valid, but redundant* arguments, which is the lowest among the analysed samples (7.84%). Finally, the highest rate of logical errors (17.65%) might be due to a combination of factors, including the complexity of the question answering task and the adopted crowd-sourcing mechanism, which prevent a thorough quality assessment.

e-SNLI. The sample includes the highest percentage of valid arguments with a total of 31.37%. However, we noticed that the complexity of the reasoning involved in e-SNLI is generally lower than WorldTree and QASC. This observation is supported by the highest percentage of *valid, but redundant* cases (31.37%), where the explanation simply repeats the content of the conclusion. This occurrs quite often for examples of lexical entailment, where the words in the conclusion are a subset of the words in the premise. The lexical entailment instances, in fact, do not require any additional world knowledge, making any attempt of constructing an explanation redundant. Despite these characteristics, our evaluation suggests that a significant percentage of arguments are invalid (37.25%). Again, this percentage might be the results of different factors, including the errors produced by the crowd-sourcing process.

Table 6.5 reports a set of representative cases extracted from the evaluated samples. For each entailment verification class, we report an example extracted from the XGS with the highest percentage of instances in that class.

6.5 Related Work

An emerging line of research in Explanation-based NLI focuses on the creation of datasets enriched with human-annotated explanations and rationales [173]. These resources are often adopted as Explanation Gold Standards (XGSs), providing additional supervision for training and evaluating explanation-based models capable of generating natural language explanations in support of their predictions [93, 18, 153, 123].

XGSs are designed to support Natural Language Inference, asking human-annotators to transcribe the reasoning required for deriving the correct prediction [151]. Despite the popularity of these datasets, and their application for measuring explainability on tasks such as Textual Entailment [15], Multiple-choice Question Answering [179, 74, 87, 72], and other inference tasks [168, 45, 44, 10], little has been done to provide a clear understanding on the nature and the quality of the reasoning encoded in the explanations.

Previous work on explainability evaluation has mainly focused on methods for assessing the quality and faithfulness of explanations generated by deep learning models [17, 146, 93, 65, 174]. Our work is related to this research, but focuses instead on the resources on which explainable models are trained. In that sense, this chapter is more aligned to gold standard evaluation methods, which aim to design systematic approaches to qualify the content and the inference capabilities involved in mainstream

NLP benchmarks [102, 14, 137, 128, 118, 114]. However, to the best of our knowledge, none of these methods have been adopted to provide a critical assessment of human-annotated explanations present in XGSs.

6.6 Conclusion and Future Work

This chapter proposed a systematic annotation methodology to quantify the logical validity of human-annotated explanations in Explanation Gold Standards (XGSs). The application of the framework on three mainstream datasets led us to the conclusion that a majority of the explanations represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors.

The main limitation of the framework lies in the scalability of its current implementation, which is generally time consuming and requires trained semanticists. One way to improve its efficiency is to explore the adoption of supporting tools for the formalisation, such as semantic parsers and/or automatic theorem provers.

Despite the current limitations, this study offers some important pointers for future work. On the one hand, the results suggest that logical errors can be reduced by a careful design of the gold standard, such as authoring explanations with internal verification strategies or reducing the complexity of the reasoning task. On the other hand, the finding that a large percentage of curated explanations still represent incomplete arguments has a deeper implication on the nature of explanations and on what annotators perceive as a valid and complete logical argument. Therefore, we argue that future progress on the design of XGSs will depend, among other things, on a better formalisation and understanding of the inferential properties of explanations.

6.7 Scoping and Limitations

The translation from natural language into logical forms represents the main limitation and bottleneck of the proposed methodology. The translation process, in fact, is generally time-consuming, requires experts in the field, and can introduce potential errors in the annotation. While we found that modelling the annotation process through a multi-label classification task can yield an acceptable inter-annotator agreement, the complexity of the methodology still limits the number of samples that can be thoroughly analysed. Therefore, additional work is required to alleviate the translation bottleneck and scale the annotation methodology to larger corpora.

Chapter 7

Conclusion

7.1 Summary and Conclusion

In this chapter, we provide a summary of how the research questions stated in Chapter 1 have been answered across the thesis, together with the main findings and a discussion on some of the main limitations of the presented work.

- ***RQ1:*** *"What is the nature and function of an explanatory argument from an epistemological-linguistic perspective?"*

  Contribution. Chapter 2 presented an extensive study on the notion of a scientific explanation from an epistemological-linguistic perspective. Specifically, the first part of the chapter focused on surveying and summarising the main modern accounts of scientific explanation developed in Philosophy of Science, identifying the constraints that these accounts impose on *explanatory arguments*, and highlighting their essential features and function.

  Findings. The systematic survey allowed us to derive the following conclusions: (1) Explanations and predictions have a different structure. An explanation, in fact, cannot be entirely characterised in terms of *deductive-inductive arguments* or *statistical relevance* relationships. This is because predictive power, despite being a necessary property of a scientific explanation, is not a sufficient one. (2) Explanatory arguments create unification. From an epistemic perspective, the main function of an explanatory argument is to fit the explanandum into a *broader unifying pattern*. (3) Explanations possess an intrinsic causal-mechanistic nature.

From an ontic perspective, a scientific explanation must cite part of the causal history of the explanandum, fitting the event to be explained into a *causal nexus*.

Scoping and limitations. The survey focused on epistemological accounts that attempt to define an *objective* relationship between *explanandum* and *explanans*. While characterising explanatory arguments is important for a complete understanding of the concept, explanation is a broader topic that embraces different aspects not considered in the survey, such as cognitive processes, conversational acts, as well as pragmatic and contextual elements involved in humans' communication [113]. While these aspects might be relevant for the construction of Explanation-based NLI models, they were considered out-of-scope for the thesis and left as a possible focus for future work. Regarding the surveyed accounts, while some consensus on the nature and function of explanation exists, philosophers still disagree on whether the discussed features apply to all types of scientific explanations and are transferable across different fields and domains [131]. Therefore, additional work is still required to derive a complete and universally accepted account and investigate whether the considered features are suitable for a general description of explanations.

- ***RQ2:*** *"How do linguistic patterns emerge in natural language explanations?"*

Contribution. The second part of Chapter 2 presented a systematic analysis of corpora of natural language explanations in the scientific domain, adopting a mixture of qualitative and quantitative methodologies to characterise *explanatory patterns* in terms of the *Causal-Mechanical* and *Unificationist* accounts. Specifically, the aim of the study was to provide complementary insights on the nature of explanations as manifested in natural language, deriving an epistemological-linguistic grounding for the construction of Explanation-based NLI models.

Findings. The corpus analysis allowed drawing the following conclusions: (1) Natural language explanations are not limited to causes and mechanisms. While *constitutive* and *etiological* elements represent the core part of an explanation, the analysis suggests that additional knowledge categories such as *definitions*, *properties* and *taxonomic relations* play an equally important role in natural

language. (2) Even if not intentionally modelled, *unification* seems to be an emergent property of corpora of explanations in the scientific domain, manifesting as explicit inference patterns in natural language; (3) Unification is realised through a process of abstraction, which represents the inference substrate connecting the explanandum to underlying explanatory regularities.

Scoping and limitations. While relating the corpus analysis to epistemological accounts allows drawing conclusions that are generalisable to some extent, the presented quantitative methodology relies on specific features of the analysed resources. Specifically, the discussed method adopted to investigate the emergence of patterns of unification could only be applied on corpora with a reuse-oriented design such as WorldTree. With the current methodology, in fact, it is not yet clear how to possibly identify such patterns through the re-occurrence of specific facts in corpora that do not possess this property.

- ***RQ3:*** *"To what extent can explicit explanatory patterns in natural language explanations improve accuracy and alleviate semantic drift for Explanation-based NLI?"*

Contribution. Chapter 3 presented a framework for Explanation Regeneration that ranks atomic facts through the combination of two scoring functions: a *relevance* score that adopts sparse lexical features, and an *explanatory power* score that leverages explicit explanatory patterns in the form of *unification*. We performed an extensive evaluation adopting a combination of k-NN clustering and sparse Information Retrieval (IR) techniques. In Chapter 4, we proposed to integrate Abductive NLI in a *case-based reasoning* framework to leverage explicit inference patterns in the explanations. Specifically, we presented a case-based abductive NLI model that retrieves and adapts natural language explanations from training examples to construct new explanations for unseen cases and address downstream inference problems, extensively evaluating the impact of the case-based framework on commonsense and scientific NLI tasks.

Findings. The experiments presented in Chapter 3 demonstrate that the proposed method achieves results competitive with some of the existing Transformer-based models, yet being orders of magnitude faster. Moreover, we empirically

show the key role of the unification-based mechanism in improving the regeneration of many-hops explanations (6 or more facts) and explanations requiring complex inference (+12.0 Mean Average Precision). Finally, we show that the constructed explanations can support downstream question answering through Abductive NLI, improving the accuracy of a BERT baseline by up to 10% overall. Similarly, the experiments in Chapter 4 demonstrated the efficacy of the case-based framework, showing that the proposed model can be effectively integrated with different sentence encoders and downstream Transformers, achieving strong performance when compared to existing multi-hop and explanation-based approaches. Moreover, we studied the impact of the retrieve-reuse-refine paradigm on explanation generation and semantic drift, finding that the case-based framework boosts the accuracy on the most challenging hypotheses.

Scoping and limitations. The model of explanatory power presented in Chapter 3 relies on the availability of human-annotated explanations with specific features (e.g., explanatory facts reused across different hypotheses). However, these resources might not be available in real-world scenarios and are generally costly to develop. Moreover, since the explanatory power model relies on similarities measures, the model's ability to generalise might be susceptible to the incompleteness of the facts bank and the availability of representative explanations. Finally, due to the use of the indicator function, the current implementation of the model is not able to identify sentences in the facts bank that have different surface forms but same underlying meaning, preventing the ability to estimate the explanatory power of sentences that are not explicitly used in the gold explanations. Regarding the proposed case-based framework in Chapter 4, The refine process adopts some simplified assumptions to model the abstraction process required for explanation generation. This process, in fact, is performed by assuming that abstraction at the concept level translates in a correct mapping between hypotheses and central explanatory sentences. However, contextual linguistic elements should be taken into account in this process as they can affect the overall meaning of the sentence and of the specific concept being abstracted. While contextual elements are partially considered during the precedent phases and have been shown to improve accuracy, additional work is required to guarantee the correctness of the refine phase and the adopted abstraction/instantiation mechanism.

- ***RQ4:*** *"Can hybrid models based on latent and explicit representations provide*

a framework for a better accuracy-scalability trade-off in Explanation-based NLI?"

Contribution. Chapter 5 focused on bi-encoders, which allow for efficient explanation-based inference via Maximum Inner Product Search (MIPS). Specifically, the chapter presented a hybrid architecture that combines a Transformer-based bi-encoder with a sparse model of explanatory power, designed to capture explicit inference patterns in corpora of scientific explanations. The model integrates sparse and dense encoders to define a joint model of relevance and explanatory power and perform multi-hop inference in an iterative fashion, conditioning the probability of selecting a fact at time-step t on the partial explanation constructed at time-step $t-1$.

Findings. We performed an extensive evaluation focusing on the trade-off between accuracy and scalability. Specifically, the chapter presented the following conclusions: (1) The hybrid framework based on bi-encoders significantly outperforms existing sparse models, achieving performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster. (2) We study the impact of the hybridisation on semantic drift, showing that it makes the model more robust in the construction of challenging explanations requiring long reasoning chains. (3) We investigate the applicability of the hybrid model on downstream Abductive NLI without additional training, demonstrating improved accuracy and robustness when performing explanation-based inference iteratively. (4) We perform a scalability analysis by gradually expanding the adopted facts bank, showing that the proposed approach can scale to corpora of millions of facts.

Scoping and limitations. The autoregressive inference model assumes that the relevance of a fact at each time step t depends on the whole explanation sequence constructed at time $t-1$. However, while it is reasonable to construct explanations in an iterative fashion, the notion of relevance might depend only on a subpart of the partially constructed explanation. In that sense, the relevance dependencies between sentences in the explanation might actually induce a sparse graphical structure. While cross-encoders can model these dependencies through the self-attention mechanism in Transformers, the same mechanism cannot be leveraged using bi-encoders. Therefore, additional work is required

for modelling this feature of explanatory relevance, including the investigation of different forms of representations (e.g. graph-based), the adoption of sparse concatenation mechanisms during autoregressive inference, and the exploration of hybrid scalable solutions preserving self-attention such as poly-encoders [63].

- ***RQ5:*** *"Do natural language explanations in existing gold standards represent valid and complete logical arguments?"*

  Contribution. Chapter 6 provided a critical quality assessment of Explanation Gold Standards (XGSs) for NLI in terms of their logical inference properties. Specifically, we presented a systematic annotation methodology that, by translating natural language explanations into corresponding logical forms, induces a set of recurring logical violations which can then be used for quantifying quality and logical consistency in the annotated explanations. We validated the methodology on three contemporary and mainstream reference XGSs for NLI.

  Findings. Through the presented study we derived the conclusion that most of the human-annotated explanations in the analysed samples represent logically invalid arguments, ranging from being incomplete to containing clearly identifiable logical errors. More fundamentally, the paper reveals that a majority of the explanations present in XGSs contain one or more major logical fallacies, while appearing to be coherent and complete on the surface.

  Scoping and limitations. The translation from natural language into logical forms represents the main limitation and bottleneck of the proposed methodology. The translation process, in fact, is generally time-consuming, requires experts in the field, and can introduce potential errors in the annotation. While we found that modelling the annotation process as a multi-label classification task can yield an acceptable inter-annotator agreement, the complexity of the methodology still limits the number of samples that can be thoroughly analysed. Therefore, additional work is required to alleviate the translation bottleneck and scale the annotation methodology to larger corpora.

- ***RQ0:*** *"Can specific epistemological-linguistic aspects of scientific explanations inform the construction of more accurate and scalable Explanation-based NLI models?"*

To summarise, this thesis investigated the notion of a scientific explanation from an epistemological-linguistic perspective to inform the development of Explanation-based NLI models. Specifically, after reviewing and clarifying the nature and function of explanatory arguments in science, we focused on specific linguistic aspects related to *unification* and the emergence of *explanatory patterns* in natural language. This allowed us to define a computational model of *explanatory power* that can be flexibly integrated in Explanation-based NLI architectures for downstream inference tasks. Under certain assumptions, such as the availability of resources such as WorldTree [72], we found that the proposed model can improve robustness in challenging multi-hop inference settings and offer a viable mechanism for a better trade-off between explanation quality and computational efficiency. Additional work is required to investigate and model additional aspects highlighted in the epistemological accounts, such as the intrinsic causal-mechanistic and contrastive nature of explanations, and explore their impact for improving model creation and evaluation methodologies in Explanation-based NLI.

7.2 Opportunities for Future Research

Here, we present a list of possible opportunities for future work:

- **Causal and mechanistic inference with natural language.** The study performed in Chapter 2 demonstrated that explanations possess an intrinsic causal-mechanistic nature. However, the empirical part of the thesis mainly focused on explicit patterns related to explanatory unification without directly modelling features connected to causal inference. Since causality constitutes a fundamental requirement for explanation, it is crucial that future work investigates this aspect, possibly exploring the integration of formal tools for causal reasoning [119] with explicit and latent semantic representations for Explanation-based NLI.

- **Analogy and abstraction.** Throughout the thesis we have empirically demonstrated the impact of explicit explanatory patterns in the form of unification. The reuse of inference patterns from similar cases, in particular, is intrinsically related to analogical reasoning [149, 134]. This form of reasoning has been identified in cognitive science and AI as a crucial mechanism enabling the formation of new concepts and explanations as well as generalisation to novel and unexpected

situations [115, 109]. While this thesis explored the reuse of sentence-level patterns through analogical transfer, future work can potentially investigate the impact of patterns at a higher level of abstraction, leveraging high-level structural similarities between hypotheses and explanations [58]. Further exploring abstractive mechanisms for analogical reasoning, in fact, could potentially lead to alleviating some of the limitations of the current model of explanatory power, such as the sensitivity to the explicit reuse of specific explanatory sentences and the incompleteness of explanation-based corpora.

- **Hybrid neuro-symbolic approaches.** The thesis demonstrated the crucial role of hybrid architectures in finding a better trade-off between accuracy and scalability in Explanation-based NLI. Future work could additionally explore this line of research focusing on richer symbolic representations [48]. The integration between neural and symbolic approaches, in fact, might contribute to the development of more efficient and accurate architectures able to perform robust multi-hop reasoning and deal with the abstractive properties of natural language explanations.

- **Impact of explanations on learning and generalisation.** In this thesis we partially explored the impact of explanations on neural models, showing that they can generally improve the performance of Transformer-based architectures on downstream NLI tasks. However, additional work is required to understand the extent to which natural language explanations can provide a way to improve generalisation, reducing bias and spurious shortcuts and favouring the learning of the underlying reasoning strategies [97, 143].

- **Moving towards inference on real-world scientific text.** An important part of future work for Explanation-based NLI in the scientific domain will be to build and evaluate architectures that can operate on real-world scientific text, moving beyond the controlled environment represented by existing benchmarks. This step would require the overcoming of some of the simplified assumptions adopted in this work for modelling multi-hop inference, as well as the development of novel evaluation methodologies.

- **Evaluation methodologies for Explanation-based NLI.** Finally, from the findings presented in Chapter 2 and 6, it is evident that further research is still required for the creation of evaluation methodologies for natural language explanations.

In particular, Chapter 2 shows that the evaluation of explanations cannot be reduced to deductive inference and entailment properties, but rather constitutes a multi-dimensional problem. In that sense, future work might consider additional properties in the evaluation such as unification and causal-mechanistic inference. Explanatory unification, in particular, might represent a feature that can support both benchmark creation and novel evaluation metrics, providing a way to build more efficient and scalable solutions for the construction of explanation-centred corpora through recurring inference patterns.

7.3 Ethical Implications

The ability to generate explanations supporting a model's decisions is becoming a fundamental requirement for the application of AI systems in real-world scenarios [12]. This requirement is particularly important in contexts with high social and economic impact, where the decisions made by an AI system should be transparent and understandable by humans [62, 164]. The importance of explainability is attested by regulations (e.g., General Data Protection Regulation (GDPR)[1]) and recent research efforts attempting to improve fairness, accountability, and transparency in AI [98, 64, 2, 164]. In that regard, the research presented in this thesis is in line with these efforts and demonstrates that progress in this direction is possible while preserving efficiency and accuracy in downstream applications.

However, as demonstrated in our experiments, Explanation-Based NLI systems, in their current state, can still generate wrong or spurious explanations, and arrive at the final predictions for the wrong reasons. These results lead to some ethical implications on the applicability of Explanation-Based NLI in real-world scenarios. Firstly, since explainability can increase trust in a model's decisions, there is a risk of the end-user becoming overconfident in a model's reasoning capabilities. For instance, just because an NLI model was able to generate a correct explanation in the past, exhibiting seemingly human-like reasoning capabilities, the user could blindly trust future predictions without carefully verifying the generated explanations. Secondly, if not equipped with the right expertise and background knowledge, the end-user could be persuaded in accepting wrong explanations as correct. Therefore, given the current state-of-the-art in the field, it is important that the deployment of Explanation-Based NLI systems is still accompanied by a careful inspection of the reasoning behind a

[1] https://gdpr.eu/

particular decision and integrated into a collaborative framework with domain experts that can understand and interpret the quality of the generated explanations.

Bibliography

[1] *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press, 1994. DOI: 10.1017/CBO9780511530128.

[2] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. "Fairness, accountability, transparency in AI at scale: Lessons from national programs". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 690–690.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. "Neural module networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 39–48.

[4] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. "Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering". In: *International Conference on Learning Representations*. 2019.

[5] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. "Generating Fact Checking Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7352–7364.

[6] Pratyay Banerjee. "ASU at TextGraphs 2019 Shared Task: Explanation ReGeneration using Language Models and Iterative Re-Ranking". In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 78–84.

[7] Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. "Careful Selection of Knowledge to Solve Open Book Question Answering". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 6120–6129.

[8] Sorin Bangu. “Scientific explanation and understanding: unificationism reconsidered”. *European Journal for Philosophy of Science* 7.1 (2017), pp. 103–126.

[9] William Bechtel and Adele Abrahamsen. “Explanation: A mechanist alternative”. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36.2 (2005), pp. 421–441.

[10] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. “Abductive Commonsense Reasoning”. *arXiv preprint arXiv:1908.05739* (2019).

[11] Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. “Genericskb: A knowledge base of generic statements”. *arXiv preprint arXiv:2005.00660* (2020).

[12] Or Biran and Courtenay Cotton. “Explanation and justification in machine learning: A survey”. In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. 1. 2017.

[13] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. “A large annotated corpus for learning natural language inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://www.aclweb.org/anthology/D15-1075.

[14] Samuel R. Bowman and George E. Dahl. *What Will it Take to Fix Benchmarking in Natural Language Understanding?* 2021. arXiv: 2104.02145 [cs.CL].

[15] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. “e-snli: Natural language inference with natural language explanations”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9539–9549.

[16] Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. “Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4157–4165. DOI:

10.18653/v1/2020.acl-main.382. URL: https://aclanthology.org/2020.acl-main.382.

[17] Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. "Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4157–4165. DOI: 10.18653/v1/2020.acl-main.382. URL: https://www.aclweb.org/anthology/2020.acl-main.382.

[18] Ruben Cartuyvels, Graham Spinks, and Marie Francine Moens. "Autoregressive Reasoning over Chains of Facts with Transformers". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 6916–6930.

[19] Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. "Autoregressive Reasoning over Chains of Facts with Transformers". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6916–6930. DOI: 10.18653/v1/2020.coling-main.610. URL: https://www.aclweb.org/anthology/2020.coling-main.610.

[20] Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, and Quoc V Le. "Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension". In: *International Conference on Learning Representations*. 2019.

[21] Yew Ken Chia, Sam Witteveen, and Martin Andrews. "Red Dragon AI at TextGraphs 2019 Shared Task: Language Model Assisted Explanation Generation". In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 85–89.

[22] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. "Think you have solved question answering? try arc, the ai2 reasoning challenge". *arXiv preprint arXiv:1803.05457* (2018).

[23] R. Clarke. "HANSON, N. R. -Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science". *Mind* 69.n/a (1960), p. 267.

[24] Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. “A Study of Automatic Metrics for the Evaluation of Natural Language Explanations”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2376–2387. DOI: 10.18653/v1/2021.eacl-main.202. URL: https://aclanthology.org/2021.eacl-main.202.

[25] Carl Craver and James Tabery. “Mechanisms in science” (2015).

[26] Carl F Craver. “Role functions, mechanisms, and hierarchy”. *Philosophy of science* 68.1 (2001), pp. 53–74.

[27] Carl F Craver and William Bechtel. “Top-down causation without top-down causes”. *Biology & Philosophy* 22.4 (2007), pp. 547–563.

[28] Carl F Craver and Lindley Darden. *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press, 2013.

[29] Carl F Craver and David M Kaplan. “Are more details better? On the norms of completeness for mechanistic explanations”. *The British Journal for the Philosophy of Science* 71.1 (2020), pp. 287–319.

[30] Jennifer D’Souza, Isaiah Onando Mulang, and Sören Auer. “Team SVMrank: Leveraging Feature-rich Support Vector Machines for Ranking Explanations to Elementary Science Questions”. In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 90–100.

[31] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. “Explaining Answers with Entailment Trees”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 7358–7370.

[32] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. “A Survey of the State of Explainable AI for Natural Language Processing”. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 2020, pp. 447–459.

[33] Rajarshi Das, Ameya Godbole, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. “A Simple Approach to Case-Based Reasoning in Knowledge Bases”. *arXiv preprint arXiv:2006.14198* (2020).

[34] Rajarshi Das, Ameya Godbole, Manzil Zaheer, Shehzaad Dhuliawala, and Andrew McCallum. “Chains-of-Reasoning at TextGraphs 2019 Shared Task: Reasoning over Chains of Facts for Explainable Multi-hop Inference”. In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 101–117.

[35] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay-Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. “Case-based Reasoning for Natural Language Queries over Knowledge Bases”. *arXiv preprint arXiv:2104.08762* (2021).

[36] Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. “Retrieval, reuse, revision and retention in case-based reasoning”. *The Knowledge Engineering Review* 20.3 (2005), pp. 215–240.

[37] Dorottya Demszky, Kelvin Guu, and Percy Liang. “Transforming question answering datasets into natural language inference datasets”. *arXiv preprint arXiv:1809.02922* (2018).

[38] David Deutsch. *The beginning of infinity: Explanations that transform the world*. Penguin UK, 2011.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. 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)*. 2019, pp. 4171–4186.

[40] Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W Cohen. “Differentiable Reasoning over a Virtual Knowledge Base”. In: *International Conference on Learning Representations*. 2019.

[41] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey”. In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.

[42] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. “DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs”. 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)*. 2019, pp. 2368–2378.

[43] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. “Hierarchical Graph Network for Multi-hop Question Answering”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 8823–8838.

[44] Deborah Ferreira and André Freitas. “Natural Language Premise Selection: Finding Supporting Statements for Mathematical Text”. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020, pp. 2175–2182.

[45] Deborah Ferreira and André Freitas. “Premise selection in natural language mathematical texts”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7365–7374.

[46] Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. “Higher-order lexical semantic models for non-factoid answer reranking”. *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 197–210.

[47] Michael Friedman. “Explanation and scientific understanding”. *The Journal of Philosophy* 71.1 (1974), pp. 5–19.

[48] Artur d’Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. “Neural-symbolic learning and reasoning: contributions and challenges”. In: *2015 AAAI Spring Symposium Series*. 2015.

[49] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. “Shortcut learning in deep neural networks”. *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[50] Stuart Glennan. “Rethinking mechanistic explanation”. *Philosophy of science* 69.S3 (2002), S342–S353.

[51] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. “A survey of methods for explaining black box models”. *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[52] R. Hadsell, S. Chopra, and Y. LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 2. 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[53] Robert James Hankinson. *Cause and explanation in ancient Greek thought*. Oxford University Press, 2001.

[54] Gilbert H. Harman. “The Inference to the Best Explanation”. *The Philosophical Review* 74.1 (1965), pp. 88–95. ISSN: 00318108, 15581470. URL: http://www.jstor.org/stable/2183532 (visited on 09/27/2022).

[55] Carl G Hempel and Paul Oppenheim. “Studies in the Logic of Explanation”. *Philosophy of science* 15.2 (1948), pp. 135–175.

[56] Carl Gustav Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press, 1965.

[57] Germund Hesslow. “The problem of causal selection”. *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), pp. 11–32.

[58] Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. “Learning to Make Analogies by Contrasting Abstract Relational Structure”. In: *International Conference on Learning Representations*. 2018.

[59] Denis J Hilton. “Conversational processes and causal explanation.” *Psychological Bulletin* 107.1 (1990), p. 65.

[60] Christopher Read Hitchcock. “Salmon on explanatory relevance”. *Philosophy of Science* 62.2 (1995), pp. 304–320.

[61] Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. “Interpretation as abduction”. *Artificial Intelligence* 63.1 (1993), pp. 69–142. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(93)90015-4. URL: https://www.sciencedirect.com/science/article/pii/0004370293900154.

[62] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. “What do we need to build explainable AI systems for the medical domain?” *arXiv preprint arXiv:1712.09923* (2017).

[63] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. “Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring”. In: *International Conference on Learning Representations*. 2019.

[64] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. “Towards accountability for machine learning datasets: Practices from software engineering and infrastructure”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 560–575.

[65] Sarthak Jain and Byron C. Wallace. “Attention is not Explanation”. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://www.aclweb.org/anthology/N19-1357.

[66] Peter Jansen. “Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering?” In: *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*. 2018, pp. 12–17.

[67] Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. “What’s in an explanation? characterizing knowledge and inference requirements for elementary science exams”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 2956–2965.

[68] Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. “Framing QA as building and ranking intersentence answer justifications”. *Computational Linguistics* 43.2 (2017), pp. 407–449.

[69] Peter Jansen, Kelly J Smith, Dan Moreno, and Huitzilin Ortiz. "On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 7529–7542.

[70] Peter Jansen, Mihai Surdeanu, and Peter Clark. "Discourse complements lexical semantics for non-factoid answer reranking". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 977–986.

[71] Peter Jansen and Dmitry Ustalov. "TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration". In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 63–77.

[72] Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. "WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[73] Peter A Jansen. "A Study of Automatically Acquiring Explanatory Inference Patterns from Corpora of Explanations: Lessons from Elementary Science Exams". In: *6th Workshop on Automated Knowledge Base Construction (AKBC 2017)*. 2017.

[74] Harsh Jhamtani and Peter Clark. "Learning to Explain: Datasets and Models for Identifying Valid Reasoning Chains in Multihop Question-Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 137–150.

[75] Yichen Jiang and Mohit Bansal. "Self-assembling modular networks for interpretable multi-hop reasoning". *arXiv preprint arXiv:1909.05803* (2019).

[76] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs". *IEEE Transactions on Big Data* (2019), pp. 1–1. DOI: 10.1109/TBDATA.2019.2921572.

[77] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://aclanthology.org/2020.emnlp-main.550.

[78] Nora Kassner and Hinrich Schütze. "BERT-kNN: Adding a kNN Search Component to Pretrained Language Models for Better QA". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3424–3430. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.307.

[79] Frank C Keil. "Explanation and understanding". *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.

[80] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. "Nearest neighbor machine translation". *arXiv preprint arXiv:2010.00710* (2020).

[81] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. "Generalization through Memorization: Nearest Neighbor Language Models". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=HklBjCEKvH.

[82] Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. "On the capabilities and limitations of reasoning for natural language understanding". *arXiv preprint arXiv:1901.02522* (2019).

[83] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. "Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 252–262.

[84] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. "Question answering via integer programming over semi-structured knowledge". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2016, pp. 1145–1152.

[85] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. "Question answering as global reasoning over semantic abstractions". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[86] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. "QASC: A Dataset for Question Answering via Sentence Composition". *arXiv preprint arXiv:1910.11473* (2019).

[87] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. "QASC: A Dataset for Question Answering via Sentence Composition". *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8082–8090. DOI: 10.1609/aaai.v34i05.6319. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6319.

[88] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. "QASC: A Dataset for Question Answering via Sentence Composition". *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8082–8090. DOI: 10.1609/aaai.v34i05.6319. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6319.

[89] Tushar Khot, Ashish Sabharwal, and Peter Clark. "Answering Complex Questions Using Open Information Extraction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 311–316.

[90] Philip Kitcher. "Explanatory unification". *Philosophy of science* 48.4 (1981), pp. 507–531.

[91] Philip Kitcher. "Explanatory unification and the causal structure of the world" (1989).

[92] Janet Kolodner. *Case-based reasoning*. Morgan Kaufmann, 2014.

[93] Sawan Kumar and Partha Talukdar. "NILE : Natural Language Inference with Faithful Natural Language Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8730–8742. DOI: 10.18653/v1/2020.acl-main.771. URL: https://www.aclweb.org/anthology/2020.acl-main.771.

[94] Souvik Kundu, Tushar Khot, Ashish Sabharwal, and Peter Clark. "Exploiting Explicit Paths for Multi-hop Reading Comprehension". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2737–2747.

[95] Carmen Lacave and Francisco J Diez. “A review of explanation methods for heuristic expert systems”. *The Knowledge Engineering Review* 19.2 (2004), pp. 133–146.

[96] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 785–794.

[97] Andrew K Lampinen, Nicholas A Roy, Ishita Dasgupta, Stephanie CY Chan, Allison C Tam, James L McClelland, Chen Yan, Adam Santoro, Neil C Rabinowitz, Jane X Wang, et al. “Tell me why!–Explanations support learning of relational and causal structure”. *arXiv preprint arXiv:2112.03753* (2021).

[98] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. “Fair, transparent, and accountable algorithmic decision-making processes”. *Philosophy & Technology* 31.4 (2018), pp. 611–627.

[99] David Lewis. “Causal explanation” (1986).

[100] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[101] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. “Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021, pp. 1000–1008.

[102] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. “Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1000–1008. URL: https://www.aclweb.org/anthology/2021.eacl-main.86.

[103] Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and Bingquan Liu. “HopRetriever: Retrieve Hops over Wikipedia to Answer Complex Questions”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 15. 2021, pp. 13279–13287.

[104] Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. “Differentiable Open-Ended Commonsense Reasoning”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4611–4625. DOI: 10.18653/v1/2021.naacl-main.366. URL: https://aclanthology.org/2021.naacl-main.366.

[105] Peter Lipton. “Contrastive explanation”. *Royal Institute of Philosophy Supplements* 27 (1990), pp. 247–266.

[106] Peter Lipton. *Inference to the best explanation*. London: Routledge, 2001.

[107] Jiangming Liu and Matt Gardner. “Multi-Step Inference for Reasoning Over Paragraphs”. *arXiv preprint arXiv:2004.02995* (2020).

[108] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. *arXiv preprint arXiv:1907.11692* (2019).

[109] Tania Lombrozo. “Explanation and abductive inference”. *Oxford handbook of thinking and reasoning* (2012), pp. 260–276.

[110] Tania Lombrozo. “The structure and function of explanations”. *Trends in cognitive sciences* 10.10 (2006), pp. 464–470.

[111] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. “Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 2381–2391.

[112] Tim Miller. “Contrastive Explanation: A Structural-Model Approach”. *arXiv preprint arXiv:1811.03163* (2018).

[113] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. *Artificial Intelligence* 267 (2019), pp. 1–38.

[114] Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. “Compositional Questions Do Not Necessitate Multi-hop Reasoning”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4249–4257. DOI: 10.18653/v1/P19-1416. URL: https://www.aclweb.org/anthology/P19-1416.

[115] Melanie Mitchell. “Abstraction and analogy-making in artificial intelligence”. *arXiv preprint arXiv:2102.10717* (2021).

[116] Tom M Mitchell, Richard M Keller, and Smadar T Kedar-Cabelli. “Explanation-based generalization: A unifying view”. *Machine learning* 1.1 (1986), pp. 47–80.

[117] Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. “Learning to Attend On Essential Terms: An Enhanced Retriever-Reader Model for Open-domain Question Answering”. 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)*. 2019, pp. 335–344.

[118] Ellie Pavlick and Tom Kwiatkowski. “Inherent Disagreements in Human Textual Inferences”. *Transactions of the Association for Computational Linguistics* 7 (Nov. 2019), pp. 677–694. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00293. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00293/1879670/tacl_a_00293.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00293.

[119] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[120] Judea Pearl. “The seven tools of causal inference, with reflections on machine learning”. *Communications of the ACM* 62.3 (2019), pp. 54–60.

[121] Charles Sanders Peirce. “Harvard lectures on pragmatism”. *Collected Papers* 5 (1903), pp. 188–189.

[122] George Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. “Answering questions by learning to rank-Learning to rank by answering questions”. 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)*. 2019, pp. 2531–2540.

[123] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 4932–4942.

[124] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. New Jersey, USA. 2003, pp. 29–48.

[125] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". 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)*. 2019, pp. 3973–3983.

[126] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://www.aclweb.org/anthology/D19-1410.

[127] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[128] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. URL: https://www.aclweb.org/anthology/2020.acl-main.442.

[129] Stephen Robertson, Hugo Zaragoza, et al. "The probabilistic relevance framework: BM25 and beyond". *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[130] Wesley C Salmon. *Causality and explanation*. Oxford University Press, 1998.

[131] Wesley C Salmon. *Four decades of scientific explanation*. University of Pittsburgh press, 2006.

[132] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[133] Wesley C Salmon. *Statistical explanation and statistical relevance*. Vol. 69. University of Pittsburgh Pre, 1971.

[134] Roger C Schank, Alex Kass, and Christopher K Riesbeck. *Inside case-based explanation*. Psychology Press, 2014.

[135] RP Schank. *Explanation patterns: Understanding mechanically and creatively*. Psychology Press, 2013.

[136] Jutta Schickore. "Scientific Discovery". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University, 2018.

[137] Viktor Schlegel, Marco Valentino, André Freitas, Goran Nenadic, and Riza Theresa Batista-Navarro. "A Framework for Evaluation of Machine Reading Comprehension Gold Standards". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020, pp. 5359–5369.

[138] Gerhard Schurz. "Explanation as unification". *Synthese* (1999), pp. 95–114.

[139] Vivian Dos Santos Silva, Siegfried Handschuh, and André Freitas. "Recognizing and Justifying Text Entailment Through Distributional Navigation on Definition Graphs." In: *AAAI*. 2018, pp. 4913–4920.

[140] Vivian S Silva, André Freitas, and Siegfried Handschuh. "Exploring knowledge graphs in an interpretable composite approach for text entailment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 7023–7030.

[141] Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. "UnNatural Language Inference". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 7329–7346.

[142] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. "Explanation in case-based reasoning–perspectives and goals". *Artificial Intelligence Review* 24.2 (2005), pp. 109–143.

[143] Joe Stacey, Yonatan Belinkov, and Marek Rei. “Supervising model attention with human explanations for robust natural language inference”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10. 2022, pp. 11349–11357.

[144] Armins Stepanjans and André Freitas. “Identifying and Explaining Discriminative Attributes”. 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)*. 2019, pp. 4313–4322.

[145] Michael Strevens. “The causal and unification approaches to explanation unified—causally”. *Noûs* 38.1 (2004), pp. 154–176.

[146] Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. “Obtaining Faithful Interpretations from Compositional Neural Networks”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5594–5608. DOI: 10.18653/v1/2020.acl-main.495. URL: https://www.aclweb.org/anthology/2020.acl-main.495.

[147] Chenhao Tan. “On the Diversity and Limits of Human Explanations”. *arXiv preprint arXiv:2106.11988* (2021).

[148] Paul Thagard. “Analogy, explanation, and education”. *Journal of Research in Science Teaching* 29.6 (1992), pp. 537–544.

[149] Paul Thagard and Abninder Litt. “Models of scientific explanation”. *The Cambridge Handbook of Computational Psychology* (2008), pp. 549–564.

[150] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. “Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Oct. 2020. URL: https://arxiv.org/abs/2010.08240.

[151] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. “A Survey on Explainability in Machine Reading Comprehension”. *arXiv preprint arXiv:2010.00389* (2020).

[152] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. “Explainable Inference Over Grounding-Abstract Chains for Science Questions”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1–12. DOI: 10.18653/v1/2021.findings-acl.1. URL: https://aclanthology.org/2021.findings-acl.1.

[153] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. “Explanationlp: Abductive reasoning for explainable science question answering”. *arXiv preprint arXiv:2010.13128* (2020).

[154] Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. “TextGraphs 2021 Shared Task on Multi-Hop Inference for Explanation Regeneration”. In: *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*. Mexico City, Mexico: Association for Computational Linguistics, June 2021, pp. 156–165. DOI: 10.18653/v1/2021.textgraphs-1.17. URL: https://aclanthology.org/2021.textgraphs-1.17.

[155] Mokanarangan Thayaparan, Marco Valentino, Viktor Schlegel, and André Freitas. “Identifying Supporting Facts for Multi-hop Question Answering with Document Graph Networks”. In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019, pp. 42–51.

[156] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. “Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9073–9080.

[157] Marco Valentino and André Freitas. *Scientific Explanation and Natural Language: A Unified Epistemological-Linguistic Perspective for Explainable AI*. 2022. URL: https://arxiv.org/abs/2205.01809.

[158] Marco Valentino, Ian Pratt-Hartmann, and André Freitas. “Do Natural Language Explanations Represent Valid Logical Arguments? Verifying Entailment in Explainable NLI Gold Standards”. In: *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. 2021, pp. 76–86.

[159] Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. “Hybrid autoregressive inference for scalable multi-hop explanation regeneration”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10. 2022, pp. 11403–11411.

[160] Marco Valentino, Mokanarangan Thayaparan, and André Freitas. “Case-based Abductive Natural Language Inference”. *arXiv e-prints* (2020), arXiv–2009.

[161] Marco Valentino, Mokanarangan Thayaparan, and André Freitas. “Unification-based Reconstruction of Multi-hop Explanations for Science Questions”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 200–211. URL: https://www.aclweb.org/anthology/2021.eacl-main.15.

[162] Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980.

[163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[164] Michael Veale, Max Van Kleek, and Reuben Binns. “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174014. URL: https://doi.org/10.1145/3173574.3174014.

[165] Douglas Walton. “A new dialectical theory of explanation”. *Philosophical Explorations* 7.1 (2004), pp. 71–89.

[166] Douglas Walton. “Abductive, presumptive and plausible arguments”. *Informal Logic* 21.2 (2001).

[167] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *International Conference on Learning Representations*. 2018.

[168] Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. “SemEval-2020 Task 4: Commonsense Validation and Explanation”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 307–321. URL: https://www.aclweb.org/anthology/2020.semeval-1.39.

[169] Erik Weber, Jeroen Van Bouwel, and Leen De Vreese. *Scientific explanation*. Springer, 2013.

[170] Leon Weber, Pasquale Minervini, Jannes Munchmeyer, Ulf Leser, and Tim Rocktäschel. “Nlprolog: Reasoning with weak unification for question answering in natural language”. *arXiv preprint arXiv:1906.06187* (2019).

[171] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. “Constructing datasets for multi-hop reading comprehension across documents”. *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 287–302.

[172] Michael R Wick and William B Thompson. “Reconstructive expert system explanation”. *Artificial Intelligence* 54.1-2 (1992), pp. 33–70.

[173] Sarah Wiegreffe and Ana Marasović. “Teach Me to Explain: A Review of Datasets for Explainable NLP”. *arXiv preprint arXiv:2102.12060* (2021).

[174] Sarah Wiegreffe and Yuval Pinter. “Attention is not not Explanation”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://www.aclweb.org/anthology/D19-1002.

[175] Joseph J Williams and Tania Lombrozo. “Explanation and prior knowledge interact to guide learning”. *Cognitive Psychology* 66.1 (2013), pp. 55–84.

[176] James Woodward. *Making things happen: A theory of causal explanation*. Oxford University Press, 2005.

[177] James Woodward. “The causal mechanical model of explanation” (1989).

[178] James Woodward and Lauren Ross. “Scientific Explanation”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021.

[179] Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. “WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 5456–5473.

[180] Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. “WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference”. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5456–5473. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.671.

[181] Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. “Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=EMHoBG0avc1.

[182] Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav, Harish Tayyar Madabushi, Oyvind Tafjord, and Peter Clark. “Multi-class Hierarchical Question Classification for Multiple Choice Science Exams”. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5370–5382. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.661.

[183] Weiwen Xu, Yang Deng, Huihui Zhang, Deng Cai, and Wai Lam. “Exploiting Reasoning Chains for Multi-hop Science Question Answering”. *arXiv preprint arXiv:2109.02905* (2021).

[184] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. “Alignment over heterogeneous embeddings for question answering”. 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)*. 2019, pp. 2681–2691.

[185] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. “Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering”. 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)*. 2019, pp. 2578–2589.

[186] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. “Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4514–4525.

[187] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 2369–2380.

[188] Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. “KGˆ 2: Learning to Reason Science Exam Questions with Contextual Knowledge Graph Embeddings”. *arXiv preprint arXiv:1805.12393* (2018).

[189] Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. “Multi-Step Reasoning Over Unstructured Text with Beam Dense Retrieval”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4635–4641. DOI: 10.18653/v1/2021.naacl-main.368. URL: https://aclanthology.org/2021.naacl-main.368.

Appendix A

Unification-based Inference

A.1 Hyperparameters tuning

The hyperparameters of the model have been tuned for the optimisation of the MAP score on the dev-set. Here, we report the values adopted for the experiments described in the chapter. The Unification-based Reconstruction adopts two hyperparameters. Specifically, λ is the weight assigned to the relevance score in equation 3.1, while k is the number of similar hypotheses to consider for the calculation of the unification score (equation 3.2). The values adopted for these parameters are as follows:

1. $\lambda = 0.83$ $(1 - \lambda = 0.17)$
2. $k = 100$

A.2 BERT model

For question answering we adopt a BERT$_{BASE}$ model. The model is implemented using PyTorch (`https://pytorch.org/`) and Hugging Face (`https://huggingface.co/`) and fine-tuned using 4 Tesla 16GB V100 GPUs for 10 epochs in total with batch size 32. The final hyperparameters adopted for BERT are as follows:

- gradient_accumulation_steps = 1
- learning_rate = 5e-5
- weight_decay = 0.0
- adam_epsilon = 1e-8

- warmup_steps = 0
- max_grad_norm = 1.0

We experimented with different learning rates [1e-5, 5e-5, 1e-6] and number of explanation sentences [3, 5, 10] and found the results to be statistically significant ($p < 0.05$). Specifically, we performed a paired t-test analysis comparing the mean accuracy of BERT achieved with the RS + PW and the RS model as explanation retrievers (see Table 3.4 in Chapter 3), averaging across the aforementioned hyperparameters.

A.3 Data and code

The experiments are carried out on the TextGraphs 2019 version (`https://github.com/umanlp/tg2019task`) of the Worldtree corpus. The full dataset can be downloaded at the following URL: `http://cognitiveai.org/dist/worldtree_corpus_textgraphs2019sharedtask_withgraphvis.zip`.

The code to reproduce the experiments described in the chapter is available at the following URL: `https://github.com/ai-systems/unification_reconstruction_explanations`

Appendix B

Case-based Abductive NLI

B.1 Hyperparameters tuning

The hyperparameters of CB-ANLI have been tuned to maximise the accuracy in downstream question answering on the WorldTree dev-set. Here, we report the best values adopted for the experiments described in the chapter.

CB-ANLI BM25:

1. $\lambda = 0.83$
2. $K = 20$

CB-ANLI Sentence-BERT:

1. $\lambda = 0.87$
2. $K = 40$

For the implementation of Sentence-BERT we adopt the following package `https://pypi.org/project/sentence-transformers/` considering the *bert-large-nli-stsb-mean-tokens* model.

B.2 Concepts Extraction

The concepts in facts and hypotheses are extracted using WordNet with NLTK: `https://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html`. Specifically, given

a sentence, we define a concept as a maximal sequence of words that corresponds to a valid synset in WordNet. This allows us to consider multi-words expressions such as "living thing".

B.3 Transformers Setup

For the implementation of the Transformer model, we fine-tuned RoBERTa (*roberta-large*) for binary classification (*bc*) to predict a set of scores $S = \{s_1, s_2, ..., s_n\}$ for each candidate hypothesis in $H = \{h_1, h_2, ..., h_n\}$. The model receives as input an hypothesis h_i along with the explanation E_i for h_i. The model is optimised via cross-entropy loss to predict 1 for the correct hypothesis and 0 for the alternative hypotheses:

$$bc(\texttt{[CLS]} \mid\mid h_i \mid\mid \texttt{[SEP]} \mid\mid E_i) = s_i \tag{B.1}$$

The binary classifier is a linear layer operating on the final hidden state encoded in the `[CLS]` token. To answer the question q, the module selects the candidate answer c_a associated to the hypothesis with the highest score – i.e. $a = \text{argmax}_i\, s_i$. The model is implemented using Hugging Face (`https://huggingface.com/`) and fine-tuned using 4 Tesla V100 GPUs for 8 epochs in total. We adopted the following hyperparameters:

- batch size = 16
- learning_rate = 1e-5
- gradient_accumulation_steps = 1
- weight_decay = 0.0
- adam_epsilon = 1e-8
- warmup_steps = 0
- max_grad_norm = 1.0

We experimented with different learning rates [1e-5, 5e-5, 1e-6] and number of central explanation sentences [1, 2, 3] and found the results to be statistically significant ($p < 0.05$). Specifically, we performed a paired t-test analysis comparing the mean accuracy of RoBERTa achieved with the CB-ANLI models and the baselines as explanation retrievers (see Table 4.3 in Chapter 4), averaging across the aforementioned hyperparameters.

B.4 Source Code

The complete code adopted in the experiments is available at the following URL: https://github.com/ai-systems/case_based_anli.

B.5 Data

The WorldTree corpus can be downloaded at the following url: http://cognitiveai.org/dist/worldtree_corpus_textgraphs2019sharedtask_withgraphvis.zip. The AI2 Reasoning Challenge dataset can be downloaded at the following URL https://allenai.org/data/arc.

Appendix C

Hybrid Autoregressive Inference

C.1 Dense Encoder

For the implementation of the dense encoder $d(\cdot)$ we adopt Sentence-BERT, whose package can be found at the following URL: `https://pypi.org/project/sentence-transformers/`. Specifically, we implement the bi-encoder with a *bert-base-uncased* model, adopting a mean-pooling operation to obtain fixed sized sentence embeddings and contrastive loss for training. We release the trained model adopted in the experiments at the following URL: `https://drive.google.com/file/d/1iz38q8EIYZdO9U7mAMVz1qUprU8jmEwI/view`.

C.1.1 Training Setup

We train the model using 4 Tesla V100 GPUs for 3 epochs in total with contrastive loss, while 10% of the training data is used for warm-up. We obtained the best results for SCAR using the following hyperparameters for training:

- batch size = 16
- margin (contrastive loss) = 0.25
- learning_rate = 2e-5
- weight_decay = 0.1
- adam_epsilon = 1e-8
- max_grad_norm = 1.0

We experimented with a different number of negative examples for training [1, 3, 5, 10] and obtained the best results with 5 negative instances.

C.1.2 Faiss Index

For creating the index of dense vectors for the facts bank we use the Faiss package for Python available at the following URL: `https://pypi.org/project/faiss-gpu/`. Specifically, we adopt IndexIVFFlat.

C.2 Source Code and Data

The complete code adopted to run our experiments is available at the following URL: `https://github.com/ai-systems/hybrid_autoregressive_inference`. The WorldTree corpus can be downloaded at the following url: `http://cognitiveai.org/dist/worldtree_corpus_textgraphs2019sharedtask_withgraphvis.zip`.