Case-Based Abductive Natural Language Inference

Marco Valentino, Mokanarangan Thayaparan, André Freitas
Idiap Research Institute, Switzerland
University of Manchester, United Kingdom
`marco.valentino@idiap.ch`

Abstract

Most of the contemporary approaches for multi-hop Natural Language Inference (NLI) construct explanations considering each test case in isolation. However, this paradigm is known to suffer from semantic drift, a phenomenon that causes the construction of spurious explanations leading to wrong conclusions. In contrast, this paper proposes an abductive framework for multi-hop NLI exploring the retrieve-reuse-refine paradigm in Case-Based Reasoning (CBR). Specifically, we present *Case-Based Abductive Natural Language Inference (CB-ANLI)*, a model that addresses unseen inference problems by analogical transfer of prior explanations from similar examples. We empirically evaluate the abductive framework on commonsense and scientific question answering tasks, demonstrating that CB-ANLI can be effectively integrated with sparse and dense pre-trained encoders to improve multi-hop inference, or adopted as an evidence retriever for Transformers. Moreover, an empirical analysis of semantic drift reveals that the CBR paradigm boosts the quality of the most challenging explanations, a feature that has a direct impact on robustness and accuracy in downstream inference tasks.

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 (Thayaparan et al., 2020). In the context of Natural Language Inference (NLI), this task is often used to develop and evaluate explanation-based systems, capable of performing transparent multi-step reasoning with natural language (Wiegreffe and Marasovic, 2021; Jansen et al., 2018; Camburu et al., 2018). While multi-hop inference has been largely explored for extractive problems such as open-domain question answering (Yang et al., 2018), increasing attention is being dedicated to
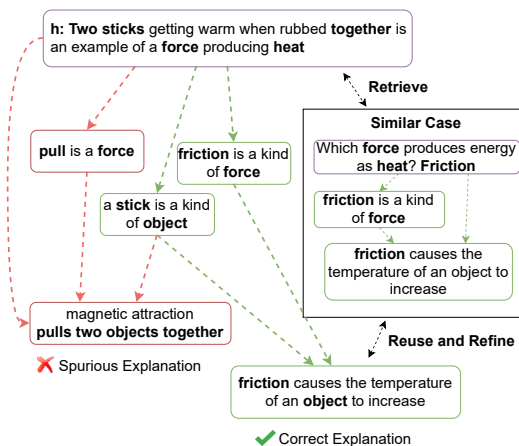

Figure 1: Performing multi-hop inference considering each case in isolation can lead to the construction of spurious explanations. In contrast, we propose the adoption of a *Case-Based Reasoning (CBR)* paradigm where the construction of new explanations is constrained by previously solved examples.

the abstractive setting, where the models are required to compose long chains of facts expressing abstract commonsense and scientific knowledge (Clark et al., 2018; Valentino et al., 2022).

In this setting, multi-hop inference is often framed as an Abductive Natural Language Inference (ANLI) problem (Bhagavatula et al., 2019), 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 (Thayaparan et al., 2021a). Existing approaches address abductive inference considering each test hypothesis in isolation, employing iterative and path-based methods (Kundu et al., 2019; Yadav et al., 2019b) or explicit constraints to guide the generation of a plausible explanation graph supporting the correct answer (Khashabi et al., 2016; Khot et al., 2017).

However, this paradigm poses several challenges in the abstractive setting as:

- 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;

- 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;

- Background knowledge sources contain a large amount of irrelevant facts overlapping with the hypothesis, a feature that can lead to the generation of spurious explanations.

Consequently, existing approaches often suffer from a phenomenon known as *semantic drift* (Khashabi et al., 2019) – i.e., the tendency of composing incorrect reasoning chains leading to wrong conclusions as the number of required inference steps increases. The example in Figure 1 illustrates some of these challenges.

In contrast with the dominant paradigm, we propose to integrate Abductive Natural Language Inference in a Case-Based Reasoning (CBR) framework (Schank et al., 2014; Das et al., 2021). CBR systems operate under 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 (Schank, 2013; De Mantaras et al., 2005). In the context of multi-hop 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:

- Similar natural language hypotheses tend to require similar explanations;

- Abstract facts tend to express general explanatory knowledge about underlying regularities, being frequently reused to explain a large variety of phenomena;

- Prior solutions can explicitly help constrain the search space, reducing the risk of composing spurious inference chains.

To investigate these hypotheses, we present a *Case-Based Abductive NLI (CB-ANLI)* model that retrieves and adapts natural language explanations

from training examples to construct new explanations for unseen cases and address downstream inference problems. Specifically, this paper provides the following main contributions:

- To the best of our knowledge, we are the first to propose an end-to-end case-based abductive framework for multi-hop and explanation-based NLI;

- We empirically demonstrate the efficacy of the CB-ANLI on commonsense and scientific reasoning tasks, 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;

- We investigate the impact of the retrieve-reuse-refine paradigm on semantic drift, and how this affects accuracy and robustness of the model. Our results show that the case-based framework boosts the quality of the explanations for the most challenging problems, resulting in improved downstream inference performance.

2 Case-based Abductive NLI

For a given set of alternative natural language hypotheses $H = \{h_1, h_2, \ldots, h_n\}$, the goal of Abductive NLI 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 and score an explanation for h_i, we adopt a Case-Based Reasoning (CBR) paradigm composed of three major phases, retrieve-reuse-refine, which can be summarised as follows (see Fig. 2):

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
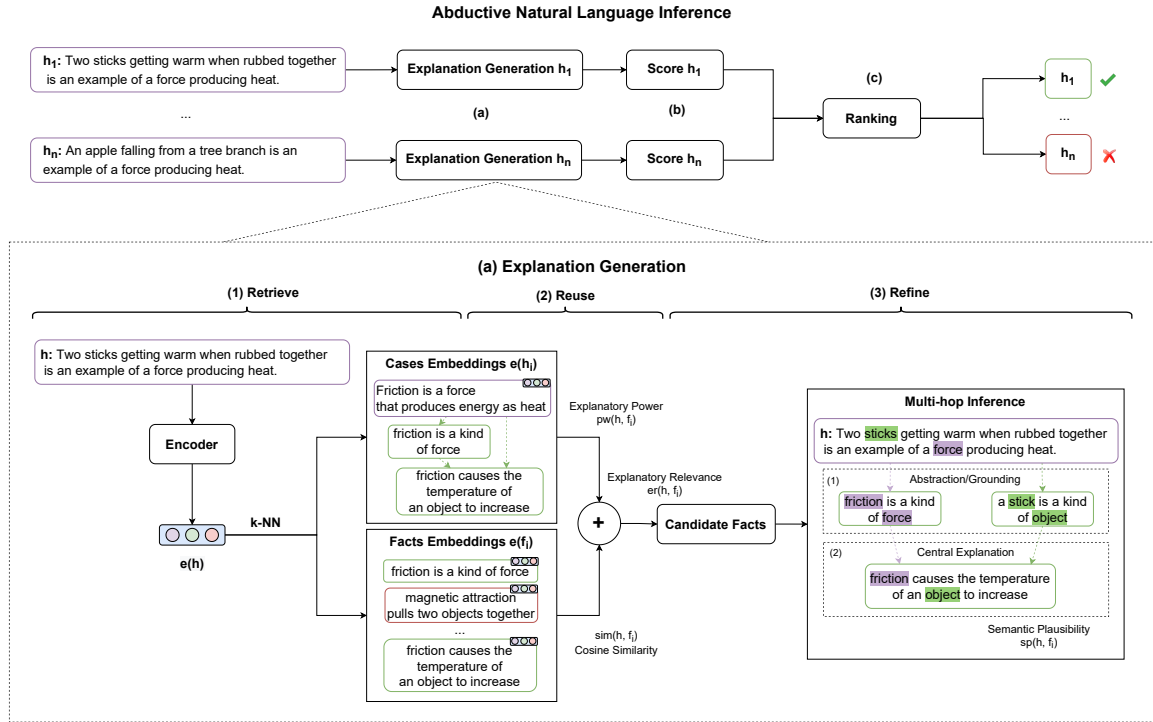
Figure 2: Overview of the proposed CBR framework. We adopt a retrieve-reuse-refine paradigm to construct and score explanations for a set of mutually exclusive hypotheses (a) and address NLI tasks via abductive inference by selecting the hypothesis supported by the best explanation.

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 sentence appears 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 inference chains to estimante the plausibility of the candidate explanatory sentences.

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

3 Explanation Generation

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 $E_i = \{f_1, \ldots, f_n\}$ form the Facts Embeddings.

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 (Robertson et al., 2009)) and a pre-trained dense encoder (Sentence-BERT (Reimers and Gurevych, 2019)) adopting a search index for efficient retrieval (IndexIVFFlat in FAISS (Johnson et al., 2019)). We adopt the WorldTree corpus (Jansen et al., 2018) as background knowledge (additional details in Section 5).

3.2 Reuse

Previous work has shown that explanatory facts expressing underlying regularities tend to create explanatory patterns across similar hypotheses (Valentino and Freitas, 2022; Valentino et al., 2021, 2022). Following this line of research, we conjecture that explanations from similar cases can be reused to constraining the search space for unseen hypotheses and improve downstream NLI.

Specifically, 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:

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

$$\mathbb{1}(f_i, h_k) = \begin{cases} 1 & \text{if } f_i \in E_k \\ 0 & \text{if } f_i \notin E_k \end{cases} \quad (2)$$

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 $e(f_i)$:

$$er(h, f_i) = \lambda \cdot sim(e(h), e(f_i)) + (1 - \lambda) \cdot pw(h, f_i) \quad (3)$$

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

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 bags 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 $CP(s_i) \cap CP(s_j)$.

To link the hypotheses to potentially abstract explanatory sentences, we construct 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 facts (Jansen et al., 2018; Thayaparan et al., 2021a) (e.g., linking *stick* to *object* and *friction* to *force* in Figure 2).

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

$$sp(h, f_i) = \frac{\sum_{cp_j \in CP(h)} path(cp_j, f_i)}{|CP(h)|} \quad (4)$$

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

4 Abductive Inference

To derive the final explanation for a given hypothesis while conditioning on previously solved cases, we sum the *explanatory relevance* computed during the reuse phase with the *semantic plausibility* computed during the refine phase, pruning the graph considering only the top n central explanatory sentences and their linked grounding nodes (Fig. 2.3).

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.

Model	Overall	Easy	Challenge	Explanation
Sparse Retrieval				
BM25 ($k = 1$) (Clark et al., 2018)	41.21	44.96	32.99	yes
BM25 ($k = 2$)	43.62	48.54	32.73	
BM25 ($k = 3$)	45.87	50.76	35.05	
Dense Retrieval				
S-BERT ($k = 1$) (Reimers and Gurevych, 2019)	44.91	50.99	31.44	yes
S-BERT ($k = 2$)	45.79	51.45	33.25	
S-BERT ($k = 3$)	44.51	49.82	32.73	
Path-based				
PathNet (Kundu et al., 2019)	41.50	43.32	**36.42**	yes
Transformers				
BERT-large (Devlin et al., 2019)	46.19	52.62	31.96	no
RoBERTa-large (Liu et al., 2019)	**50.20**	**57.04**	35.05	no
Case-based Abductive NLI				
CB-ANLI BM25 ($n = 1$)	52.13	56.34	42.78	yes
CB-ANLI BM25 ($n = 2$)	<u>55.17</u>	60.42	<u>**43.56**</u>	
CB-ANLI BM25 ($n = 3$)	52.69	58.56	39.69	
CB-ANLI S-BERT ($n = 1$)	54.45	<u>**61.23**</u>	39.43	yes
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 1: Accuracy on WorldTree (test-set) for *easy* and *challenge* questions. The parameter n corresponds to the number of central explanatory sentences considered by the models to compute the scores for each hypothesis.

5 Empirical Evaluation

Experimental Setup. We evaluate the Case-based Abductive NLI (CB-ANLI) framework on WorldTree (Jansen et al., 2018) and AI2 Reasoning Challenge (ARC) (Clark et al., 2018), 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 (Demszky et al., 2018).

The knowledge bases required for the inference are populated using the WorldTree corpus (Jansen et al., 2018). 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 ≈ 20), requiring challenging multi-hop inference to be generated. We store the individual facts for deriving the *Facts Embeddings* and consider the training questions ($\approx 1K$) and their explanations as the set of previously solved cases (*Cases Embeddings*). For the refine phase, we dynamically extract the concepts in facts and hypotheses using WordNet (Miller, 1995) with

NLTK[1]. Additional details are described in the appendix[2].

Sentene Encoders. We evaluate CB-ANLI using sparse and dense sentence encoders without additional training. The sparse version adopts BM25 vectors (Robertson et al., 2009), while the dense version employs Sentence-BERT (large) (Reimers and Gurevych, 2019; Thakur et al., 2020).

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 categories of approaches: *Retrieval Solvers*, *Path-based Solvers*, and *Transformers*. The results in terms of question answering accuracy are reported in Table 1.

Retrieval Solvers. We employ stand-alone BM25 and Sentence-BERT (large) as sparse and dense retrieval solvers (Clark et al., 2018). Given

---

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

[2] Code available at the following url: https://github.com/ai-systems/case_based_anli

RoBERTa + Retriever	Over.	Easy	Chal.
None	50.20	57.04	35.05
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 2: Results for RoBERTa large fine-tuned on WorldTree and augmented with different explanation retrieval models.

an hypothesis h, the solvers retrieve the top 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 (Kundu et al., 2019) as a multi-hop inference baseline. This model constructs inference paths connecting question and candidate answer, and 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 a BERT large (Devlin et al., 2019) and a RoBERTa large (Liu et al., 2019) baseline fine-tuned on the multiple-choice question answering task. We observe that on WorldTree the proposed approach is competitive with both RoBERTa and BERT (up to $\approx 5\%$ and $\approx 9\%$ improvement respectively).

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

Explainable Models	Accuracy
TupleInf (Khot et al., 2017)	23.83
TableILP (Khashabi et al., 2016)	26.97
DGEM (Clark et al., 2018)	27.11
KG^2 (Zhang et al., 2018)	31.70
Unsupervised AHE (Yadav et al., 2019a)	33.87
Supervised AHE (Yadav et al., 2019a)	34.47
ET-RR (Ni et al., 2019)	36.61
ExplanationLP (Thayaparan et al., 2021a)	40.21
AutoROCC (Yadav et al., 2019b)	41.24
Attentive Ranker (Pirtoaca et al., 2019)	**44.72**
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 3: Performance on the AI2 Reasoning Challenge (ARC). We compare CB-ANLI with published explainable approaches that are fine-tuned only on ARC.

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 answer. Table 2 reports the accuracy achieved with RoBERTa large when adopting CB-ANLI and stand-alone baselines as evidence retrievers. In general, we observe that evidence retrieval plays an important role for improving the performance of RoBERTa, and that the use of CB-ANLI can generate useful explanations for inference in combination with downstream language models.

5.2 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) (Clark et al., 2018). Here, we keep the same configuration of hyperparameters. Table 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, supporting generalisation on new
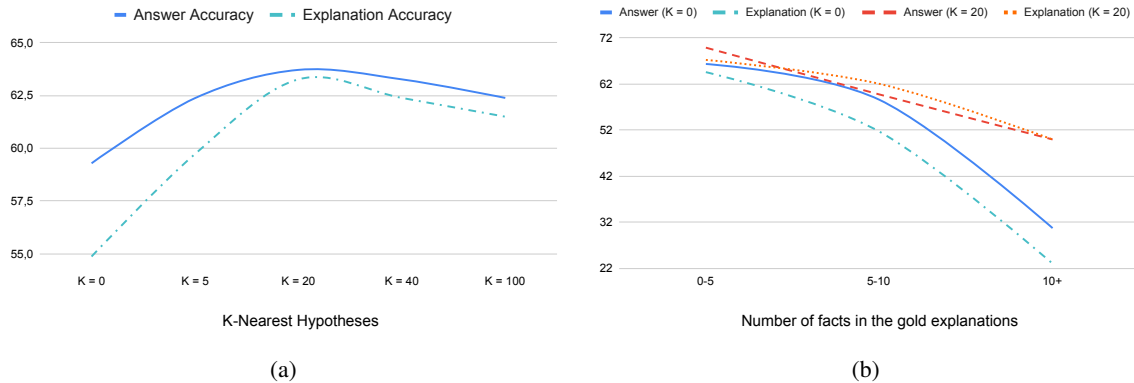
Figure 3: Impact of the case-based framework on semantic drift. K represents the number of similar cases considered for computing the explanatory power (Worldree dev-set).

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 explainable approaches that are fine-tuned only on ARC, without additional pre-training on related datasets (e.g. OpenBookQA (Mihaylov et al., 2018), RACE (Lai et al., 2017)). The results show that CB-ANLI S-BERT is third in the ranking, outperforming explanation-based systems based on Integer Linear Programming (ILP) (Khot et al., 2017; Khashabi et al., 2016) and pre-trained embeddings (Yadav et al., 2019a). At the same time, CB-ANLI obtains competitive results when compared with most of the fine-tuned neural approaches, including ET-RR (Ni et al., 2019). Moreover, when combined with RoBERTa, CB-ANLI achieves the best results among the considered approaches, improving on AutoROCC (Yadav et al., 2019b) and Attentive Ranker (Pirtoaca et al., 2019).

5.3 Ablation Study

We carried out an ablation study to investigate the impact of the CBR framework on downstream inference performance. To this end, we consider different versions of CB-ANLI by removing the impact of the reuse and refine phase. For the first, we remove the explanatory power term in equation 3. For the latter, we simply skip the refine phase ignoring the explanation graph construction and the semantic plausibility score to filter the central explanatory facts. The results of the study, reported in Table 4, demonstrate the key role of each phase to achieve the final inference performance.

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: Ablation Study on WorldTree (test-set) by removing the impact of the reuse and refine phases.

5.4 Impact on Semantic Drift

In this section, we investigate the impact of the CBR paradigm on semantic drift and how this affects the results on downstream reasoning tasks. To this end, we measure the performance of CB-ANLI when retrieving a different number K of previously solved hypotheses (notice that when $K = 0$ the model is equivalent to a non-case-based method). To evaluate the quality of the generated explanations, we leverage the WorldTree corpus as a gold standard, computing the *explanation accuracy* as the percentage of the best central explanations retrieved by the model that are part of the gold explanations in WorldTree. Since the explanations in the test-set are not publicly available, we perform this analysis on the dev-set.

Figure 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 improved explanation generation capabilities. Specifically, by conditioning the inference on an increasing number of similar hy-

1562

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 5: Examples of explanations constructed by CB-ANLI. The underlined choices represent the correct answers. *Accurate* indicates whether the central fact **(bold)** is part of the gold explanation in the WorldTree corpus.

potheses, CB-ANLI is able to construct more accurate explanations, a feature that has a direct impact on downstream inference performance in question answering.

Figure 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 (Jansen and Ustalov, 2019). 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.

5.5 Faithfulness and Error Analysis

Finally, we present an analysis of 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 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. The example number 4 shows the case

in which an accurate explanation is not sufficient to discriminate the correct answer. We found this cases to occur for a total of 31.71% of incorrect answers. Finally, the last row describes the situation in which wrong answers are caused by inaccurate or spurious explanations (for a total of 68.29% of the wrong answers). This analysis demonstrates the interpretability and faithfulness of the framework, showing that its behaviour can be typically traced back to the quality of the generated explanations.

6 Related Work

Multi-hop inference for abstractive tasks is challenging as the general structure of the explanations cannot be derived from the surface form of the NLI 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 (Fried et al., 2015; Khashabi et al., 2019).

Existing approaches frame multi-hop inference as the problem of building an optimal graph, conditioned on a set of semantic constraints (Khashabi et al., 2018; Khot et al., 2017; Jansen et al., 2017; Khashabi et al., 2016; Thayaparan et al., 2021a), or adopting iterative methods, using sparse or dense encoding mechanisms (Yadav et al., 2019a,b; Pirtoaca et al., 2019; Kundu et al., 2019). Our model is related to previous work that leverages annotated explanations to reduce semantic drift (Xie et al., 2020; Jansen et al., 2018). However, this line of work is still limited to explanation regeneration (Jansen and Ustalov, 2019; Cartuyvels et al., 2020; Valentino et al., 2021, 2022; Thayaparan et al., 2021b), while the applicability of these re-

sources for downstream multi-hop NLI problems is yet to be explored. In this paper, we move a step forward, exploring how the impact of annotated explanations on semantic drift translates in improved downstream performance.

Our approach is related to previous work on Case-Based Reasoning (CBR) (Schank et al., 2014; Schank, 2013; De Mantaras et al., 2005). Similar to the retrieve-reuse-refine paradigm adopted in CBR systems, 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. Khandelwal et al. (2020b,a) adopt k-NN search to retrieve similar training examples and improve pre-trained language models and machine translation without additional training. Similarly, Das et al. (2021, 2020) propose a CBR framework for knowledge base reasoning, while Kassner and Schütze (2020) reuse similar cases to improve BERT (Devlin et al., 2019) on cloze-style QA. To the best of our knowledge, this is the first application of Case-Based Reasoning for explanation-based multi-hop Natural Language Inference (NLI).

The work presented in this paper is related to hybrid neuro-symbolic approaches for inference with natural language (Liu et al., 2020; Minervini et al., 2020; Jiang and Bansal, 2019; Chen et al., 2019; Dua et al., 2019; Xu et al., 2021; Weber et al., 2019). In this context, most of the existing approaches combine neural models with symbolic programs or reasoning modules. For instance, Jiang and Bansal (2019) propose the adoption of a Neural Module Network (Andreas et al., 2016) for multi-hop question answering by designing four atomic neural modules (Find, Relocate, Compare, NoOp). Weber et al. (2019) propose a methodology to perform multi-hop inference using a Prolog prover via the integration of sentence encoders and a weak unification mechanism. Differently from the methodology discussed in this paper, previous neuro-symbolic approaches have been generally applied to extractive tasks, where the inference steps (and, therefore, the explanation's structure) can be derived from a direct decomposition of the questions (Thayaparan et al., 2020).

7 Limitations

The adopted model of explanatory power 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 similarity measures and indicator functions, the model's ability to generalise might be sensitive to the incompleteness of the knowledge bases and the availability of representative explanations. We believe these limitations can be potentially alleviated by exploring the role of more abstract sentence representations within the CBR paradigm (Bergmann and Wilke, 1996).

In the current implementation of CB-ANLI, the refine phase adopts specific 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 can still affect the overall meaning of the specific concept being abstracted, inducing the inclusion of spurious links between sentences. While contextual elements are considered during the precedent phases through the use of contextualised embeddings and similar cases, additional work is still required to guarantee the correctness of the abstraction process.

8 Conclusion

This paper presented CB-ANLI, a model that integrates multi-hop and Case-Based Reasoning (CBR) in a unified framework. We demonstrated the efficacy of the framework in complex abstractive and multi-hop NLI tasks. We believe this work can open new lines of research on hybrid neuro-symbolic models for explanation-based NLI, and plan to investigate the efficacy of the framework on architectures that adopt richer symbolic representations in combination with neural models, further exploring the role of abstraction in Case-Based Reasoning for improving robustness, generalisation, and explainability in NLI.

Acknowledgements

References

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

Ralph Bergmann and Wolfgang Wilke. 1996. On the role of abstraction in case-based reasoning. In *European Workshop on Advances in Case-Based Reasoning*, pages 28–43. Springer.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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

Ruben Cartuyvels, Graham Spinks, and Marie Francine Moens. 2020. Autoregressive reasoning over chains of facts with transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6916–6930.

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

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

Rajarshi Das, Ameya Godbole, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2020. A simple approach to case-based reasoning in knowledge bases. In *Automated Knowledge Base Construction*.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611.

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. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, 20(3):215–240.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. 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)*, pages 2368–2378.

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

Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. 2017. Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*, 43(2):407–449.

Peter Jansen and Dmitry Ustalov. 2019. 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)*, pages 63–77.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. 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)*.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. 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)*, pages 4474–4484.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, pages 1–1.

Nora Kassner and Hinrich Schütze. 2020. BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In *Findings of the Association for Computational Linguistics: EMNLP*

2020, pages 3424–3430, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020a. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020b. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

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

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

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

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 311–316.

Souvik Kundu, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019. Exploiting explicit paths for multi-hop reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2737–2747.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Jiangming Liu, Matt Gardner, Shay B Cohen, and Mirella Lapata. 2020. Multi-step inference for reasoning over paragraphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3040–3050.

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. 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, pages 2381–2391.

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

Pasquale Minervini, Matko Bošnjak, Tim Rocktäschel, Sebastian Riedel, and Edward Grefenstette. 2020. Differentiable reasoning on large knowledge bases and natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5182–5190.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2019. 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)*, pages 335–344.

George Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. 2019. 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)*, pages 2531–2540.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3973–3983.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

RP Schank. 2013. *Explanation patterns: Understanding mechanically and creatively*. Psychology Press.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. 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*.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021a. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12, Online. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. 2021b. 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)*, pages 156–165.

Marco Valentino and André Freitas. 2022. Scientific explanation and natural language: A unified epistemological-linguistic perspective for explainable ai. *arXiv preprint arXiv:2205.01809*.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. 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*, pages 200–211.

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. Nlprolog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. 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*, pages 5456–5473.

Weiwen Xu, Yang Deng, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Exploiting reasoning chains for multi-hop science question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1143–1156.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019a. 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)*, pages 2681–2691.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019b. 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)*, pages 2578–2589.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. 2018. Kgˆ 2: Learning to reason science exam questions with contextual knowledge graph embeddings. *arXiv preprint arXiv:1805.12393*.

A   Hyperparameters

We adopted the following hyperparameters for CB-ANLI:

CB-ANLI BM25:

1. $\lambda = 0.83$

2. $K = 200$

CB-ANLI S-BERT:

1. $\lambda = 0.97$

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   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"* that frequently occur in the scientific domain.

C   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:

$$bc(\texttt{[CLS]} \parallel h_i \parallel \texttt{[SEP]} \parallel E_i) = s_i \quad (5)$$

The binary classifier is a linear layer operating on the final hidden state encoded in the $\texttt{[CLS]}$ 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

D   Source Code

The code adopted in the experiments is available at the following URL: https://github.com/ai-systems/case_based_anli.

E   Data

The WorldTree corpus adopted in the experiments can be downloaded at the following url: http://cognitiveai.org/dist/worldtree_corpus_textgraphs2019sharedtask_withgraphvis.zip. The AI2 Reasoning Challenge (ARC) dataset can be downloaded at the following URL: https://allenai.org/data/arc. For the experiments on ARC, we adopted WorldTree V2 as our background knowledge: http://cognitiveai.org/explanationbank/