External Knowledge-Driven Argument Mining: Leveraging Attention-Enhanced Multi-Network Models

Debela Gemechu and **Chris Reed**
Centre for Argument Technology (ARG-tech)
University of Dundee
Dundee DD1 4HN, United Kingdom
{d.t.z.gemechu,c.a.reed}@arg.tech

Abstract

Argument mining (AM) involves the identification of argument relations (AR) between Argumentative Discourse Units (ADUs). The essence of ARs among ADUs is context-dependent and lies in maintaining a coherent flow of ideas, often centered around the relations between discussed entities, topics, themes or concepts. However, these relations are not always explicitly stated; rather, inferred from implicit chains of reasoning connecting the concepts addressed in the ADUs. While humans can infer such background knowledge, machines face challenges when the contextual cues are not explicitly provided. This paper leverages external resources, including WordNet, ConceptNet, and Wikipedia to identify semantic paths (knowledge paths) connecting the concepts discussed in the ADUs to obtain the implicit chains of reasoning. To effectively leverage these paths for AR prediction, we propose attention-based Multi-Network architectures. Various architecture are evaluated on the external resources, and the Wikipedia based configuration attains F-scores of 0.85, 0.84, 0.70, and 0.87, respectively, on four diverse datasets, showing strong performance over the baselines.

1 Introduction

Argument mining involves identifying the argumentative structure within a text. It includes segmenting arguments into Argumentative Discourse Units (ADUs) (Peldszus and Stede, 2015a), distinguishing argumentative units from non-argumentative ones, classifying ADUs, labeling argument relation (AR) between ADUs, and identifying argument schemes (Persing and Ng, 2016; Stab and Gurevych, 2017; Lawrence and Reed, 2020). This study focuses on classifying the AR between ADUs into three categories: Inference (RA) (when one ADU supports the other), Conflict (CA) (when one ADU attacks the other), and None (when there is no AR).

The nature of AR is inherently context-dependent (Potash et al., 2017; Habernal et al., 2017; Choi and Lee, 2018; Rinott et al., 2015), relying on maintaining a coherent flow of interconnected ideas. This cohesion is often centered around the connections between the discussed entities, topics, themes or concepts, commonly referred to as **Local coherence** (Foltz et al., 1998; Marcu, 2000). Local coherence facilitates smooth idea transitions between ADUs by recognising inherent regularities in entity distribution. Similarly, other entity-based theories of discourse (Givón, 1987; Prince, 1981) and **Centering Theory** (Grosz et al., 1995) propose that these regularities contribute to the coherence of discourse by guiding the organisation of ideas around salient entities. Following a similar framework, aspect-based argument mining techniques use the relationships between the concepts discussed in ADUs, to identify argument structures (Misra et al., 2017; Dragoni et al., 2018; Gemechu and Reed, 2019; Trautmann, 2020). Yet, the contexts required to link these concepts are not always explicit and are often inferred from background knowledge.

Pre-trained large language models (LLMs) have transformed NLP, moving from traditional feature engineering to data-driven approaches. Studies indicate that these models implicitly capture various types of knowledge, including relational, commonsense, and structural linguistic knowledge, within their parameters (Petroni et al., 2019; Goldberg, 2019; Safavi and Koutra, 2021; AlKhamissi et al., 2022). While excelling in various NLP tasks, their ability to encode the necessary background knowledge for identifying ARs remains uncertain (Kassner and Schütze, 2019). For example, Polu et al. (2022) revealed their limitations in chaining multiple steps of complex logical reasoning, while Merrill et al. (2021) demonstrated they fail to comprehend the semantics behind commonsense reasoning tasks. This limitation is critical in AR iden-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3688–3709
November 12-16, 2024 ©2024 Association for Computational Linguistics

tification, as linking ADUs relies on the implicit chain of reasoning, often inferred from the chain of relations between the concepts discussed in the ADUs. This highlights the need for supplementary contextual information from external sources to establish these connections.

Consider the ADUs from the 2016 presidential election debate corpus (Visser et al., 2019) in Table 1. Identifying the AR between (1) and (2) relies on recognising the relationship between "NAFTA agreement" and "USA", whereas for (4) and (5), it requires understanding "building electric grid" is an "economic activity". While these connections are straightforward for human experts, computers face challenges as such interconnections are often implicitly inferred. For example, the AR between (3) and (4) is direct as the relation between the concepts mentioned in the respective ADUs can be obtained from an ontology (Miller, 1995; Speer et al., 2017) ("Electric grid; grid" is directly related to "power; electrical power" in WordNet (Miller, 1995)) or by comparing their embeddings (Pilehvar et al., 2013; Le and Mikolov, 2014; Reimers and Gurevych, 2019). However, identifying the AR between (4) and (5) is challenging since the path linking "electric grid" to "economic activity" is missing in existing knowledge resources including WordNet (Miller, 1995) or ConceptNet (Speer et al., 2017) or DBpedia.[1] However, the concepts are indirectly linked in Wikipedia through a chain of concepts interlinked using a set of semantic relation types: "economic activity" *involves* "innovation" which *constitutes* developing "clean energy" *transmitted by* "electric grid". This study aims to identify and leverage the chain of such semantic relations between the concepts, to capture implicit referential information between ADUs (Asher and Lascarides, 2003) and use it for AR prediction.

No	ADUs
1	[USA]$_C$ [is in deep trouble]$_{OC}$
2	[NAFTA agreement]$_C$ [is defective]$_{OC}$
3	[We]$_C$ [can have]$_{OC}$ [clean energy]$_A$
4	[We]$_C$ [can build]$_{OC}$ a new modern [electric grid]$_A$
5	[This]$_C$ [is a lot of]$_{OC}$ new [economic activity]$_A$

Table 1: Examples from 2016 presidential election debate corpus (Visser et al., 2019) to illustrate the relation between the functional components of ADUs. *C* represents the theme of the sentence, *A* represents the aspects specialising the theme, while the opinion on *C* is represented by *OC*.

Leveraging knowledge from external resources has been shown to improve performance in AM (Kobbe et al., 2019; Botschen et al., 2018; Fromm et al., 2019; Plenz et al., 2023) and related tasks, such as semantic plausibility (Wang et al., 2018), identifying inferences (Chen et al., 2017), and determining entailment (Glockner et al., 2018). However, existing studies on AR prediction exclusively utilise structured knowledge bases and overlook semi-structured resources like Wikipedia, which contains over 6,805,837 articles (as of April 1, 2024), offering richer connections through hyperlinks embedded within articles. Moreover, these methods rely on entities, events, and factual information sourced from structured databases, limiting their applicability to specific domains. In contrast, using generic semantic relation types that encode AR ensures adaptability across domains (refer to Table 7 for examples of such relation types). Furthermore, they lack effective method for integrating the external information into model architectures, relying instead on conventional feature engineering techniques. For instance, Kobbe et al. (2019) leverage features derived from graph representations of the resources, including the inter-concept distances within the graph. Similarly, Plenz et al. (2023) employ semantic similarity to determine the relevance of external knowledge, in conjunction with traditional features derived from the graph representation of the resources.

In this paper, we propose traversing Wikipedia, WordNet, and ConceptNet to find semantic paths linking concepts mentioned in ADU pairs. ARs between ADUs are identified by leveraging these paths using attention-based Multi-Network architectures. To establish a benchmark, we evaluate LLMs across various configurations, comparing the knowledge obtained from external resources with that inherent in LLMs. The evaluation demonstrated that integrating external resources consistently enhances performance, showing strong performance over the baselines and comparison approaches. Additionally, we assess the effectiveness of the attention-based Multi-Network architecture in leveraging external knowledge, demonstrating its superiority over the standard linear classification baseline. The contribution of this paper is fourfold: (a) the utilisation of both structured and semi-structured external resources for AR prediction, (b) architecture for effectively leveraging external knowledge, (c) features adaptable across domains, and (d) the state-of-the-art performance.

[1] https://wiki.dbpedia.org/

2 Related Works

In the literature, AM has been approached using various configurations, including dependency parsing (Peldszus and Stede, 2015b), discourse parsing (Muller et al., 2012), sequence tagging (Eger et al., 2017; Mayer et al., 2020), and sequence classification configurations (Reimers et al., 2019; Ruiz-Dolz et al., 2021; Mayer et al., 2020). Various works tackle specific AM tasks. Some focus exclusively on argument segmentation (Chernodub et al., 2019; Ajjour et al., 2017), while others start with segmented data and focus solely on AR identification (Potash et al., 2016; Gemechu and Reed, 2019; Ruiz-Dolz et al., 2021). Potash et al. (2016) train an encoder-decoder (Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2014) to identify AR. Gemechu and Reed (2019) decompose ADUs into fine-grained components and use classifiers to predict AR based on the relations between these components. Chakrabarty et al. (2020) identify argument components and ARs within both inter-turn and intra-turn interactions in dialogues. They classify ARs as a binary prediction, determining only the presence of a relation without specifying its type. Their findings indicate that using distant-labeled data and integrating discourse relations from Rhetorical Structure Theory (Mann and Thompson, 1988) improve performance.

End-to-end AM approaches address multiple AM tasks, simultaneously. Persing and Ng (2016) and Stab and Gurevych (2017) adopt a pipeline architecture and train separate models for each subtask to then utilise an Integer Linear Programming (ILP) model to encode global constraints. Eger et al. (2017) propose a neural end-to-end approach, framing the task in various configurations including dependency parsing and token-based sequence tagging. They also employ a multi-task setup to leverage the dependencies between AM tasks, including component identification and AR prediction. Their best-performing configuration achieves an F1-score of 0.51 for AR identification on the AAEC dataset. Peldszus and Stede (2016) aim to map RST trees to argumentation structures (Taboada and Mann, 2006) using sub-graph matching and an evidence graph model. They evaluate various features of their system on the AMT dataset and achieve an overall F-measure of 0.76 in identifying ARs. Similarly, Morio et al. (2022) introduce an end-to-end cross-corpus training strategy that facilitate information transfer between datasets. Mayer et al. (2020) address argument component and relation identification on a dataset comprising various disease treatments. The approach involves combining static and dynamic embeddings using various configurations of RNN and CRFs. They demonstrate the efficacy of specialised LLMs like SciBERT (Beltagy et al., 2019), highlighting their relevance in medical domain adaptations. However, most of these works rely on the information explicitly provided in the argument alone.

Recent AM works fine-tuned LLMs in sequence classification fashion (Reimers et al., 2019; Ruiz-Dolz et al., 2021). Studies show that such LLMs implicitly capture relational, commonsense, and structural linguistic knowledge (Petroni et al., 2019; Goldberg, 2019; Safavi and Koutra, 2021; AlKhamissi et al., 2022). Despite their significant performance, the ability of LLMs to encode the requisite background knowledge for identifying ARs remains uncertain, raising concerns about relying solely on LLMs for this task (Kassner and Schütze, 2019). For instance, Polu et al. (2022) exposed their limitations in complex logical reasoning, while Merrill et al. (2021) showed they struggle in comprehending the semantics of commonsense reasoning tasks.

The works most related to ours are those of Kobbe et al. (2019) and Plenz et al. (2023), as they also leverage external knowledge bases to identify AR. However, their methodologies differ significantly from ours. Firstly, they rely on structured knowledge bases with predefined relation types, while we also use semi-structured resources like Wikipedia that cover diverse relations. Furthermore, they struggle to effectively integrate external information into model architectures, relying instead on conventional feature engineering techniques that exploit structural features obtained from sub-graph extracted from external knowledge bases. For instance, Kobbe et al. (2019) use features like the frequency of relations existing between ADUs. Similarly, Plenz et al. (2023) leverage the similarity between external knowledge and ADUs to identify relevant sub-knowledge graphs and exploit the sub-graph to extract categorical features, such as the number of shared concepts between ADU pairs and the path lengths between the concepts. Additionally, the formalisation of the "concepts" used for alignment with external resources is vague, relying on arbitrary entity mentioned in the ADUs. Moreover, their approach for AR identification has not been evaluated.

3 Methodology

Our approach comprises two main stages: first, we align a pair of ADUs (premise-conclusion) with external knowledge resources and extract the relevant knowledge paths connecting them; second, we incorporate these knowledge paths into our model architecture using attention-based Multi-Networks to predict the AR between the ADUs. The following subsections provide details on the data used and the processes involved in each step.

3.1 Data

We use four corpora. The first is AAEC (Stab and Gurevych, 2017) which has a total of 402 arguments. ADUs under each argument are labelled as premise, claim or major claim. It has 147,271 tokens, 6,089 ADUs and 5335 ARs (4841 support and 497 attack).

The second corpus is the Argumentative Micro Text (AMT) (Peldszus and Stede, 2013) which is a collection of 112 short texts collected from human subjects in German translated into English. It is annotated following the argumentation structure outlined by MicroTextAnnotation. The structure consists of a central claim, and supporting ADUs. It has a total of 8007 tokens, 576 ADUs and 443 ARs (272 support and 171 attack).

The third corpus is part of the US 2016 presidential election debate corpus (US2016) (Visser et al., 2019) which is annotated based on Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011). Argument components are referred to as propositions, with the relations between them annotated as default inference for support and default conflict for attack. The corpus has a total of 15805 tokens, 1473 ADUs and 584 ARs (505 support and 79 attack).

The fourth corpus is the AbstRCT corpus (Mayer et al., 2020) which consists of abstracts extracted from the MEDLINE database. Argument components are categorised into major claim, claim, and evidence components, and the relations between them are categorised into support, attack, and partial-attack. The corpus consists of 100,253 tokens, 4,679 ADUs, and 2,634 ARs, including 344 attack relations (combining attack and partial-attack relations) and 2,290 support relations.

As described above, argument components are annotated non-uniformly across datasets, based on the underlying theoretical framework. For example, in AAEC, argument components are annotated as premises, claims, and major claims. However, in US2016, the components are not explicitly categorised, but the premise-conclusion notion can be inferred from the direction of the AR. As our current objective does not involve classifying the components or the direction of the relation, we focus on the AR existing between the components without classifying the categories of the components (into claim/conclusion/major-claim, premise/evidence).

3.2 External Knowledge Alignment and Paths Extraction

Each ADU is annotated into its four functional components, following the framework proposed by Gemechu and Reed (2019) (see Appendix A.3 for more details). These components are used to align the respective ADUs with the external resources. The functional components consist of target concepts (***C***), aspects (***A***), opinions on ***C*** (***OC***), and opinions on ***A*** (***OA***). ***C*** refers to the set of concepts related to the ADUs's topic, while ***A*** refers to the set of concepts further specifying that topic (examples provided in Table 1). In this study, we focus on ***C*** and ***A***, which represent the topics and aspects addressed by the ADUs. The statistics of these components can be found in Table 4 in the Appendix.

To extract relevant external knowledge, we align these components with two ontological resources—WordNet (Miller, 1995) and ConceptNet (Speer et al., 2017)—as well as a semi-structured resource, Wikipedia. The detailed alignment process is described in Sections 3.2.1 to 3.2.2.

3.2.1 Ontology as External Source

We traverse WordNet (Miller, 1995) and ConceptNet (Speer et al., 2017) Synset hierarchies and align the components of ADUs with the Synsets, to identify the chain (path) of Synsets that connects the components. The alignment relies on cosine similarity between the embeddings of the components and Synsets, determined by the cosine similarity threshold β. Sentence-transformer (Reimers and Gurevych, 2019) is utilised to identify the embeddings. For more details on the embeddings and similarity threshold, check Appendix A.3.2.

By treating the ontology as a graph, with Synsets as nodes and relation types as edges, we begin the search with one of the components and traverse the knowledge graphs until either the other component is found or the search depth reaches the threshold $\alpha = 5$. For more details on setting the value of

α, refer to Appendix A.3.3. If the search is successful, we concatenate the Synsets and the type of semantic relation between, otherwise return the concatenation of both components with the constant string "None" in between. We use relation types with frequency higher than *m=3* to form the paths (see Appendix A.3.4 for more information about the relation filtering process).

3.2.2 Wikipedia as External Source

We also traverse Wikipedia to identify the chain (path) of Wikipedia pages linking the functional components of ADUs. For any pair of components (e.g., C_1, A_2 or C_1, C_2 or A_1, A_2) associated with a pair of ADUs (p_1, p_2), the initial step involves aligning these components with corresponding Wikipedia pages. This alignment is achieved by computing the similarity between the embeddings (Reimers and Gurevych, 2019) of the Wikipedia page titles and the components.

Viewing Wikipedia as a graph (with pages as nodes and hyperlinks as edges), we begin a breadth-first search from the Wikipedia page of one concept (c_1), continuing until we locate the second concept (c_2) or reach a depth threshold, $\alpha = 5$. During this search, we record sentences (S) containing Wikipedia page titles of the current page (hl_1) and the hyperlinks leading to the next Wikipedia page (hl_2) along the path. These sentences contribute to the formation of a tuple: $\langle hl_1, hl_2, keywords\rangle$, where the keywords represent the semantic relation type linking hl_1 and hl_2 within the sentences.

We utilise semantic role labeling (SRL) to identify the keywords that connect hl_1 and hl_2 within the sentences (S) containing the hyperlinks. The SRL tool from AllenNLP [2] is used for this purpose. The process involves extracting subject-predicate structures that link hl_1 and hl_2 in the sentences involving the hyperlinks, followed by identifying phrases that connect them across the semantic roles assigned (see Appendix A.3.5). Top *m* most frequent relations are selected to construct the paths.

3.3 Model

We propose attention-based Multi-Network to leverage the information obtained from external resources for AR prediction (Section 3.3.1). Section 3.3.2 presents baseline models that utilise LLMs alone as sources of background knowledge.

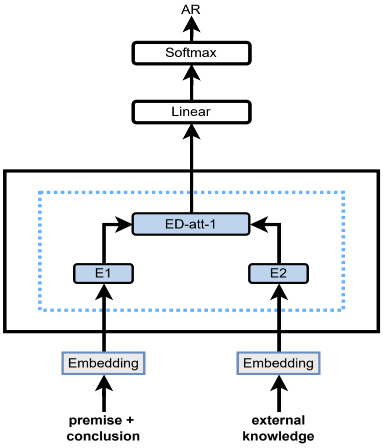

Figure 1: Siamese-networked with attention layers.

3.3.1 Attention-Based Multi-Network

We investigate two attention-based Multi-Network configurations, namely Siamese and Triplet (Schroff et al., 2015) networks, built using pre-trained LLM blocks. Initially, we utilise the Siamese network involving two sub-networks, where one sub-network encodes the concatenation of both ADUs together while the other encodes the external information. Furthermore, we examine Triplet network, which uses three sub-network to encodes each of the ADUs and the external resources separately.

Siamese Network Architecture with Attention. In this setup, given the two sub-networks (**E1** and **E2**) in Siamese network, **E1** processes the concatenation of the pair of ADUs (premise and conclusion), while **E2** handles the concatenation of the information from external resources. Cross attention layer (Vaswani et al., 2017) (**ED-att-1**) is applied to the outputs of these sub-networks for attending to the external resources relevant to the premise and conclusion (see Figure 1). Accordingly, the output of **E1**, which represents the premise and conclusion, functions as the query, while the output of **E2**, representing external knowledge, serves as the keys and values. This setup allows the premise and conclusion to query relevant external information. It employs multi-head attention (Vaswani et al., 2017) h, where each head j computes scaled dot-product attention using query $\mathbf{Q}^j$, key $\mathbf{K}^j$, and value $\mathbf{V}^j$ matrices, which are linear transformations of the input hidden state $\mathbf{h}_i$. The final attention weight $\mathbf{e}_i$ is obtained by concatenating over all attention heads. The result-

[2] https://docs.allennlp.org/v0.9.0/api/allennlp.models.semantic_role_labeler.html

ing attention weights are then multiplied with the output of **E1**, and passed through a fully connected classification layer, for AR classification. This fusion allows the model to integrate the original representations of the premise and conclusion with the extracted external information (see detailed model parameters in the Appendix A.5).

Triplet Network Architecture with Attention. In contrast to the Siamese Architecture, the Triplet Network Architecture consists of three subnetworks: **E1**, **E2**, and **E3** (see Figure 3 in Appendix 3.3.1). Sub-networks **E1** and **E2** encode the premise and conclusion, respectively, while **E3** encodes the external knowledge connecting them. Two cross-attention layers are introduced (**ED-att-1** and **ED-att-2**). **ED-att-1** focuses on the relation between the premise and conclusion, where the output of **E1** serves as queries and the output of **E2** is used as keys and values. On the other hand, **ED-att-2** attends to the external knowledge relevant to the premise and conclusion. Specifically, the output of **ED-att-1** acts as the query, while the output of **E3** is used as keys and values. Similar to the Siamese architecture, we combine the output of the two attention layers for classification. The rationale behind this approach is that **ED-att-1** encodes the relation between the premise and conclusion, while **ED-att-2** encodes the relevant external resource, enabling the model to effectively leverage both the relationship between the premise and conclusion and the relevant external knowledge for AR classification.

3.3.2 LLMs as Baseline Models

We establish LLMs without external resources as baseline models under two configurations: few-shot and fully fine-tuning configurations. We evaluate these baselines against configurations that leverage external knowledge sources to enhance the performance of LLMs.

Few-shot setup: We prompt GPT-4[3], a generative LLM, to perform two tasks: (a) predicting ARs for comparative analysis against models utilising external resources, and (b) generating paths between ADU components for comparison with models using paths derived from ontology and Wikipedia. Accordingly, GPT-4-generated paths are used as external knowledge to train the Multi-Network configuration for AR classification. This enables a direct comparison between GPT-4-generated paths and those obtained from other external knowledge sources. The experimental setup for prompting GPT-4 is provided in A.4.

Fine-tuning setup: We also fine-tuned BERT (Devlin et al., 2018) using various configurations for comparison. Initially, we use the vanilla sequence classification setup (SC$\odot$V$\oplus$bert), where the concatenation of ADUs is presented as an input. Furthermore, we fine-tune BERT within Siamese architectures, both with (SM$\odot$A$\oplus$bert) and without attention layers (SM$\odot$V$\oplus$bert). See A.1 for the details of model configuration and experimental setups.

4 Experiments

4.1 Experimental Setup

The dataset is randomly partitioned, with 70%, 10%, and 20% allocation for training, validation, and testing respectively, ensuring uniformity throughout the dataset. Refer to Table 3 in the Appendix for the breakdown of ARs accross the datasets. Results represent the average of three runs using different random seeds. Precision (P), recall (R), and F-measure (F) are computed, and macro-averaged P, R, and F are reported for the test dataset (more experimental setup provided in Appendix A.1). The datasets and code used in our experiments are publicly available.[4]

4.2 Model Configurations

The encoder blocks in both the Siamese and Triplet networks are built using BERT. The cross-attention layers use 8 heads, matching standard transformer architecture, with outputs concatenated and then passed to the feedforward and classification layers. Both the Siamese and Triplet networks use a single BERT model for all encoders, sharing identical parameters across the networks. We apply cross-entropy loss based on the final classification layer output, as the Triplet loss in the multi-network architecture is not suitable for AR prediction. We evaluate various configurations leveraging the two ontological resources (WordNet, and ConceptNet) and Wikipedia across the four datasets. These configurations encompass three Triplet network architectures: TL$\odot$A$\oplus$wp for Wikipedia, TL$\odot$A$\oplus$wn for WordNet, and TL$\odot$A$\oplus$cn for ConceptNet. Similarly, three Siamese network architectures are evaluated across these ontological resources: SM$\odot$A$\oplus$wp, SM$\odot$A$\oplus$wn, and SM$\odot$A$\oplus$cn.

[3] https://openai.com/chatgpt

[4] https://github.com/arg-tech/ExternalKnowledge-ArgMining

Configs		**AAEC**			**AMT**			**US2016**			**AbstRCT**	
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Comparison												
P2016	n/a	n/a	77	n/a	n/a	74	n/a	n/a	n/a	n/a	n/a	n/a
K2019	n/a	n/a	59	n/a	n/a	67	n/a	n/a	n/a	n/a	n/a	n/a
PS2016	n/a	n/a	n/a	n/a	n/a	76	n/a	n/a	n/a	n/a	n/a	n/a
E2017	n/a	n/a	51	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
GPT-4	63±2.0	48±2.0	55±2.0	60±2.0	47±2.0	52±2.0	58±1.0	43±2.0	50±1.0	69±3.0	58±2.0	63±2.0
GR2019	81	74	77	**88**	66	75	63	61	62	n/a	n/a	n/a
M2020	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	62	n/a	n/a	69
LLMs as KB												
SC⊙V⊕bert	78±0.3	73±0.2	75±0.1	79±0.4	67±0.1	72±0.1	56±0.4	64±0.2	60±0.2	84±0.0	82±0.0	83±0.0
SM⊙V⊕bert	77±0.1	72±0.1	74±0.1	80±0.9	65±0.2	72±0.5	55±0.2	63±0.1	59±0.1	82±0.0	82±0.0	82±0.0
SM⊙A⊕bert	80±0.2	73±0.3	76±0.2	80±0.1	68±0.3	74±0.2	57±0.1	64±0.2	60±0.1	85±0.0	83±0.0	84±0.0
No Att + Ext	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TL⊙V⊕gpt	77±2.0	84±2.0	80±2.0	74±3.0	81±2.0	77±3.0	54±4.0	76±3.0	64±3.0	72±4.0	**87±3.0**	80±4.0
SM⊙V⊕wn	84±0.0	79±0.2	81±0.1	82±0.4	73±0.3	77±0.3	62±0.1	69±0.1	65±0.1	82±0.0	82±0.0	82±0.0
SM⊙V⊕cn	83±0.3	76±0.1	80±0.2	82±0.1	72±0.2	77±0.1	61±0.2	71±0.2	66±0.2	84±0.0	85±0.1	85±0.1
SM⊙V⊕wp	82±0.2	82±0.1	82±0.1	84±0.2	76±0.3	80±0.2	63±0.3	71±0.6	67±0.3	85±0.0	85±0.0	85±0.0
TL⊙V⊕wn	83±0.1	79±0.2	81±0.1	84±0.1	75±0.1	80±0.1	61±0.0	70±0.0	65±0.0	83±0.1	82±0.1	82±0.1
TL⊙V⊕cn	83±0.1	80±0.2	82±0.1	84±0.1	76±0.1	80±0.1	61±0.1	71±0.1	66±0.0	85±0.0	85±0.0	85±0.0
TL⊙V⊕wp	84±0.0	80±0.0	82±0.0	82±0.1	76±0.1	79±0.1	64±0.1	70±0.1	67±0.1	86±0.1	85±0.0	86±0.1
Att + Ext	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TL⊙A⊕gpt	77±3.0	**84±2.0**	80±3.0	71±4.0	**85±4.0**	77±3.0	56±3.0	72±3.0	63±3.0	73±2.0	85±4.0	79±3.0
SM⊙A⊕wn	84±0.1	81±0.1	82±0.1	83±0.1	79±0.1	81±0.1	62±0.1	72±0.1	67±0.1	83±0.0	82±0.0	83±0.0
SM⊙A⊕cn	83±0.3	81±0.2	82±0.2	81±0.2	82±0.2	81±0.2	65±0.1	72±0.1	68±0.1	85±0.1	85±0.0	85±0.1
SM⊙A⊕wp	85±0.2	81±0.1	83±0.1	82±0.2	83±0.2	82±0.1	65±0.2	73±0.2	69±0.2	85±0.1	86±0.0	86±0.0
TL⊙A⊕wn	85±0.1	82±0.1	84±0.1	83±0.2	84±0.2	**84±0.1**	64±0.2	72±0.3	68±0.2	83±0.1	83±0.0	83±0.0
TL⊙A⊕cn	84±0.1	82±0.2	83±0.1	82±0.1	84±0.1	83±0.1	65±0.2	73±0.3	69±0.2	86±0.0	85±0.0	86±0.0
TL⊙A⊕wp	**86±0.1**	83±0.2	**85±0.2**	83±0.1	**85±0**	**84±0.0**	**66±0.1**	75±0.1	**70±0.1**	**87±0.1**	86±0.0	**87±0.0**

Table 2: Performance of our models and the comparison systems including, (Potash et al., 2016) (P2016), (Eger et al., 2017) (E2017), (Peldszus and Stede, 2016) (PS16), (Kobbe et al., 2019) (K2019), (OpenAI, 2023) (GPT-4), (Gemechu and Reed, 2019) (GR2019), (Mayer et al., 2020) (M2020) across the four datasets. The reported results have been averaged from 3 randomly initialised sequential runs. The table is divided into subsections: Comparison approaches; LLM-alone; non-attention with external sources; attention-based with external resources.

Furthermore, to evaluate the attention layers' impact on external resources, we compare Triplet and Siamese architectures without attention layers across the three external resources, totaling six configurations: TL⊙V⊕wp, TL⊙V⊕wn, TL⊙V⊕cn, SM⊙V⊕wp, SM⊙ V⊕wn, and SM⊙V⊕cn. Finally, we evaluate the Triplet architecture on GPT-4 generated paths (TL⊙V⊕gpt, TL⊙A⊕gpt).

4.3 Results and Discussions

The evaluation results depicted in Table 2 revealed clear trends in performance. Particularly, the influence of model architecture and the incorporation of external knowledge on AR prediction. This is evident by the performance improvement observed in configurations with such integration compared to those without.

Models incorporating external resources outperformed those lacking such integration, indicating the importance of leveraging additional knowledge sources for AR identification. This led to a notable enhancement, surpassing the baseline by over 5.4% in F-measure. For example, the Siamese architecture leveraging Wikipedia achieved an average F-measure of 80% across datasets, whereas its counterpart, lacking the external resource, achieved 74%. This finding aligns with previous research demonstrating that while LLMs tend to encode world knowledge, LLMs alone may not fully present the depth and specificity of knowledge required for certain tasks, such as AR identification involving structured and chained reasoning (Kassner and Schütze, 2019; Polu et al., 2022; Merrill et al., 2021). Likewise, models equipped with attention mechanisms consistently surpassed those without, demonstrating an average increase in F-measure of over 2% across diverse configurations. Notably, Triplet Network architecture with attention mechanism leveraging Wikipedia as an external knowledge source, attained an average F-measure of 81% across the datasets. This demonstrates a significant performance improvement in AR identification, highlighting the effectiveness of the architecture in integrating external knowledge.

We also compare our approach to other related works including Potash et al.'s (2016), Eger et al.'s (2017), Peldszus and Stede's (2016), Kobbe et al.'s (Kobbe et al., 2019), OpenAI's GPT-4 (OpenAI, 2023), Gemechu and Reed's (2019; 2023) and Mayer et al.'s (2020) work. Please note that direct comparisons with some of these works need additional contextual nuance in interpretation due to variations in task setup and complexities. For instance, the works of Eger et al. (2017) and Mayer et al. (2020) involve argument segmentation in addition to AR identification as an end-to-end task. In our case, the goal is to identify AR based on correct segments in the gold datasets. Similarly, Plenz et al. (2023) evaluate their approach on several AM tasks, including ValNov Shared Task (Heinisch et al., 2022), which involves assessing the validity and novelty of a conclusion given a premise—a task closely related to AR prediction. They report an F1 score of 70.69% for this task. As can be seen from Table 2, our approach outperforms the comparison systems, including OpenAI's GPT-4 (OpenAI, 2023) across the datasets.

Model Architecture Influence. As shown in Table 2, incorporating attention layers into Multi-Network architectures brought clear benefits. Multi-network configurations with attention mechanisms outperformed the vanilla sequence classification setup, both with and without external knowledge, achieving an average F1 gains of 6.4% and 1%, respectively. Attention-based configurations leveraging external resources consistently outperform their counterparts without attention, yielding an average F1-score improvement of 2%. The attention-based Triplet architecture outperformed their counterpart Siamese architecture, with an average performance increase of 1.2% in leveraging external knowledge. It is noteworthy that in the absence of attention and external resources, multi-network configurations (SM$\odot$V$\oplus$bert) underperform as compared to the vanilla sequence classification approach (SC$\odot$V$\oplus$bert).

This highlights the efficacy of attention-based Multi-Network architectures in leveraging external resources for AR prediction, contrasting with standard sequence classification setups. Additionally, the performance advantage of Triplet architecture over Siamese architecture can be attributed to its design, enabling each sub-network to focus on learning two levels of alignment: between the premise and conclusion, and between the external resource and the premise-conclusion pair. To explore whether the performance gap solely stems from the additional parameters in the attention layer, we introduced extra linear layers to the Multi-Network architecture (without attention layers) and observed no change in performance despite the additional layers. However, attention analysis is required to substantiate this claim.

External knowledge influence. Wikipedia-based models outperform the baselines and ontology-based models across all four datasets. The attention-based Triplet-network on Wikipedia (TL$\odot$A$\oplus$wp) achieved F-measures of 0.85, 0.84, 0.70, and 0.87 in identifying AR on AAEC, AMT, US2016, and AbstRCT, respectively. Upon analysing the paths connecting the components of ADUs, we found that 37% of concepts not present in ontological resources are connected in Wikipedia, while only 7% of concepts absent in Wikipedia are covered by ontological resources. For further details, please refer to Appendix A.3.6. This disparity highlights the richness of Wikipedia's network of hyperlinks, which connect pages using a variety of relationships, unlike ontological resources that rely on predefined, narrow sets of semantic relations.

Although exploring combinations of external databases such as WordNet and ConceptNet could offer additional insights, their contributions were marginal, with only a 7% improvement over Wikipedia's paths. This motivated our focus on Wikipedia as it provides the most comprehensive set of connections. For simplicity and clarity, we chose to highlight the most impactful results rather than explore more complex combinations. However, a combined approach across all three sources could be considered in future iterations.

Models trained on GPT-4 generated paths outperformed those without external knowledge, aligning with other works leveraging LLM-generated commonsense knowledge (Bansal et al., 2022). However, despite exhibiting higher accuracy, they still demonstrated lower precision compared to approaches utilising external knowledge sources. The observed high recall and low precision can be attributed to the models' tendency to identify unintended paths between concepts. Twenty errors were randomly selected for analysis, with two human annotators collaboratively examining the paths. Of these, 14 errors were deemed contextually irrelevant, despite the logical coherence evident in the generated paths. These paths introduce chains of thought that diverge from the original argument, as noted in previous studies that rely on LLMs for generating commonsense knowledge (Levy et al., 2022).

It is important to note that in our study, GPT serves primarily as a comparison system rather than a core external resource. The Standardized Mean Difference (SMD) shows that while GPT generally outperforms baseline models, the improvement varies across datasets. The TL$\odot$A$\oplus$GPT model surpasses the baseline SM$\odot$A$\oplus$BERT, achieving an overall SMD of 0.59. For datasets like AAEC, AMT, and US2016, the average SMD is 1.57, indicating significant enhancements in those contexts. However, in the AbstRCT dataset, GPT generated paths based configurations underperform and negatively affected overall performance. This discrepancy underscores the need for careful selection and integration of external knowledge sources to enhance model efficacy. An error analysis can be found in Appendix A.3.6.

5 Conclusion

Our exploration of various model configurations underscored the importance of external resources and multi-network architecture with attention mechanisms in AR prediction. Models augmented with external resources consistently outperform those relying solely on LLMs. This emphasises the necessity of leveraging supplementary knowledge sources to enrich LLMs for AR prediction. Furthermore, multi-network architectures with attention mechanisms, notably the attention-based Triplet Network architecture, demonstrates superiority across all configurations. Further work is required to delve deeper into attention analysis, to shed-light on its role in encouraging the model to focus in aligning the premise with the conclusion, as well as in linking the premise-conclusion pair with external knowledge. While configurations leveraging Wikipedia outperformed those using other resources, more work is required to evaluate the quality of keywords representing semantic relations between concepts identified from Wikipedia against the standard semantic relation types in ontologies. Furthermore, alternative methods for extracting these keywords should be explored.

Limitations

Although our work presents promising advancements, it also entails the following limitations.

Cross-Domain Evaluation. Robust evaluation involving cross-domain evaluation, where models are trained on one domain and evaluated on a new domain, is essential for uncovering the robustness of the proposed approaches. While our evaluation has primarily focused on specific domains or datasets, cross-domain evaluation can provide insights into the generalisability and adaptability of the models across diverse domains and real-world applications.

External Knowledge Alignment and Relation Identification. More work is required in aligning the concepts with external resources, particularly in disambiguating the senses of the Synsets and Wikipedia page titles. Our current approach relies on simple similarity measures between the embeddings of glosses of the resources and the components, which may lead to missing alignments and incorrect alignment. Improving the alignment procedure to account for semantic ambiguity and variability in external resources is crucial for enhancing the effectiveness of the proposed approach. Additionally, sophisticated techniques are needed to identify the semantic relation types existing between Wikipedia hyperlinks. Unlike ontologies, Wikipedia does not encode explicit semantic relation types between hyperlinks. Therefore, developing robust method to identify semantic relations from Wikipedia articles can improve the quality and relevance of external knowledge integration in AR prediction.

Interpretability and Explainability. The explanations provided regarding the performance of the architectures and external resources are based on the analysis of empirical results. While empirical analysis is valuable for understanding model behavior, additional techniques beyond the results themselves can provide deeper insights into model performance. Exploring techniques such as model visualisation, attention mechanisms analysis, and interpretability methods like LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) or SHAP (SHapley Additive explanations) (Lundberg and Lee, 2017) can help uncover the underlying reasons behind model decisions and configurations. Complementing empirical analysis with interpretability techniques can allow a more comprehensive understanding of model behavior.

Acknowledgements

This research is supported in part by: the Swiss National Science Foundation under grant 10001FM_200857; the Volkswagen Foundation under grant 98 543; and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract 2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorised to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Neet Kaur, and Balaji Krishnamurthy. 2022. Cose-co: Text conditioned generative commonsense contextualizer. *arXiv preprint arXiv:2206.05706*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Teresa Botschen, Daniil Sorokin, and Iryna Gurevych. 2018. Frame-and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 90–96.

Katarzyna Budzynska and Chris Reed. 2011. Whence inference. *University of Dundee Technical Report*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2020. Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2017. Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

Carlos Chesnevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The knowledge engineering review*, 21(4):293–316.

HongSeok Choi and Hyunju Lee. 2018. Gist at semeval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mauro Dragoni, Celia da Costa Pereira, Andrea GB Tettamanzi, and Serena Villata. 2018. Combining argumentation and aspect-based opinion mining: the smack system. *AI Communications*, 31(1):75–95.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. Tacam: Topic and context aware argument mining. *arXiv preprint arXiv:1906.00923*.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351.

Talmy Givón. 1987. Beyond foreground and background. *Coherence and grounding in discourse*, 11:175–188.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. 2022. Overview of the 2022 validity and novelty prediction shared task. In *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. 2019. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. 2022. Safetext: A benchmark for exploring physical safety in language models. *arXiv preprint arXiv:2210.10045*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

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

Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2017. Using summarization to discover argument facets in online ideological dialog. *arXiv preprint arXiv:1709.00662*.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 196–204.

Andreas Peldszus and Manfred Stede. 2015a. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Andreas Peldszus and Manfred Stede. 2015b. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining*, pages 103–112.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and-walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351.

Moritz Plenz, Juri Opitz, Philipp Heinisch, Philipp Cimiano, and Anette Frank. 2023. Similarity-weighted construction of contextualized commonsense knowledge graphs for knowledge-intense argumentation tasks. *arXiv preprint arXiv:2305.08495*.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2016. Here's my point: Joint pointer architecture for argument mining. *arXiv preprint arXiv:1612.08994*.

Ellen F Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*, pages 1135–1144.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.

Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Tara Safavi and Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. *arXiv preprint arXiv:2104.05837*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Maite Taboada and William Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Dietrich Trautmann. 2020. Aspect-based argument mining. *arXiv preprint arXiv:2011.00633*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, pages 1–32.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

A Appendix

We provide additional details regarding the methodology and experimental setups used in our study.

A.1 Experiment Setup

A.1.1 Training Procedure

Hyper-parameters: We employ Adam optimisation (Kingma and Ba, 2014) to minimise the cost function. The learning rate is set to $2e^{-5}$ with a batch size of 16. Categorical cross-entropy loss was used as the loss function.

Gradient Clipping: To prevent exploding gradients during training, we apply gradient clipping. We use a maximum gradient norm (`max_grad_norm`) parameter set to 1.0 to determine the threshold for gradient clipping.

Warm-up and Learning Rate Schedule: We employed a linear warm-up strategy for the learning rate. The number of warm-up steps is set to 10% of the total training steps. Following the warm-up phase, the learning rate schedule is determined by a lambda function. This function linearly increases the learning rate during the warm-up phase and decreases it linearly thereafter.

Early Stopping: We implement early stopping to prevent overfitting and to determine the optimal number of epochs. This technique involves continuously monitoring the loss and F-score on the validation set throughout training. If there is a sustained degradation in performance over consecutive epochs, training is terminated to prevent the model from being influenced by noise present in the training data.

A.1.2 Input Setup

For the baseline sequence classification configurations, we concatenate the premise to the conclusion using a special token `[SEP]`. In the Siamese architecture, one of the sub-networks takes the concatenation of the premise and conclusion based on the special token `[SEP]`, while the other takes the concatenation of the paths. The paths are concatenated using the special token `[SEP]`.

The number and length of the paths between the components of the ADUs vary, with some ADUs not involving any path at all. For ADU pairs involving a large number of paths exceeding the maximum sequence length, we concatenate the paths until the maximum sequence length is reached. In such cases, we sort the paths based on their frequency. The concatenation process starts from the most frequent paths until the maximum sequence length is reached.

A.1.3 Fully Fine-tuned Baseline LLM Configuration

For the fully fine-tunned baseline LLM configuration, we utilise the HuggingFace implementation of BERT for sequence classification (`bert-base-uncased`[5]). We experimented with two variants of BERT: `bert-base-uncased` and `bert-large-uncased`. Our experiments revealed that `bert-base-uncased` consistently provided better performance compared to `bert-large-uncased`. In the baseline Siamese architecture, each sub-network independently encodes the ADUs.

Dataset	Training		Validation		Test	
	RA	CA	RA	CA	RA	CA
AAEC	4235	411	605	59	1210	117
US2016	353	55	51	8	101	16
MTC	190	120	27	17	55	34
AbstRCT	1603	241	229	34	458	69

Table 3: Distribution of support and attack relations across the training, validation, and test splits for the datasets. RA and CA refer to terms from the AIF (Chesnevar et al., 2006), where RA stands for Rule (of inference) Application, representing a relation of support or inference, and CA stands for Conflict (scheme) Application, indicating a relation of conflict or attack.

Dataset	Total C	Total A	Unique C	Unique A
AAEC	9875	6789	5634	4356
US2016	3225	1737	1854	1566
MTC	870	589	756	470
AbstRCT	7343	6432	5554	4546

Table 4: Distribution of target concepts (***C***) and aspects (***A***) across the datasets.

A.2 External Knowledge Extraction

A.3 ADU Decomposition

To identify the functional components (***C*** and ***A***) from ADUs, we adopt a sequence labeling approach following the methodology outlined by Gemechu and Reed (2019). Unlike Gemechu and Reed (2019) method, which employs a convolutional neural network (CNN), we fine-tune BERT for token classification using their dataset annotated with the BIO sequence labeling scheme, outperforming their top-performing method by 3% and achieving a macro F-score of 0.784. We utilise the HuggingFace implementation of BERT (`bert-base-uncased` [6]). The inputs are padded to 256 maximum size. We use the train-test split in the original dataset. Training is conducted over 6 epochs, and evaluation is reported as the average performance over 3 runs of the experiment on the

[5] `https://huggingface.co/google-bert/bert-base-uncased`

[6] `https://huggingface.co/google-bert/bert-base-uncased`

test dataset. Using the fine-tuned model, we identify the functional components of ADUs, and the distribution of these components is presented in Table 4.

A.3.1 Alignment of Ontologies and Wikipedia

For aligning ontologies and Wikipedia with the components of ADUs, the cosine similarity between the embeddings of the components and the Synsets of the ontologies or the corresponding Wikipedia page title is used. Additionally, we utilise the similarity between the concepts and the gloss texts of the respective sources for disambiguating senses, for concepts involving multiple senses.

A.3.2 Similarity Threshold

We leverage embeddings derived from Sentence-transformers, particularly the *all-roberta-large-v1*[7] variant, for determining similarity. We set a similarity threshold of $\beta = 0.80$ based on experimental comparisons of similarity scores between related and unrelated text pairs in the STSB dataset.[8]

The dataset is originally annotated on a scale of 0-5 based on the degree of similarity. We transform the original 5-class labels into binary labels, where labels below 4 are considered unrelated, and labels 4 and above are deemed related. In the original annotation rubric provided by SemEval-2017 (Cer et al., 2017), label 3 indicates sentences that are roughly equivalent, but some important information differs. However, we found that this definition allows for a certain degree of looseness in similarity assessment. Consequently, to impose a stricter criterion for similarity, we decided to raise the threshold from label 3 to label 4. To this end, we calculate the similarity between the sentence pairs in the training dataset and select the threshold yielding the highest F1-score. We compute F1-scores at 20 similarity threshold points (ranging from 0 to 1 with increments of 0.05), as outlined in Algorithm 1.

A.3.3 Search Depth Threshold

To estimate the optimal depth threshold for navigating through the knowledge graphs, we employ the following procedure: we randomly select 20 pairs of concepts and initiate a complete search from one

Algorithm 1 Find Optimal Similarity Threshold

```
Require: List of sentence pairs $(s_1, s_2)$
Ensure: Threshold
  best_threshold $\leftarrow$ min_thr
  max_f_score $\leftarrow 0$
  for thr $\leftarrow$ min_thr to max_thr by thr_step do
      tp $\leftarrow 0$
      fp $\leftarrow 0$
      fn $\leftarrow 0$
      for each sentence pair $(s_1, s_2)$ in data do
          sim_score $\leftarrow$ sim score$(s_1, s_2)$
          if similarity_score $\geq$ thr then
              if pair is similar then
                  tp $\leftarrow$ tp $+ 1$
              else
                  fp $\leftarrow$ fp $+ 1$
              end if
          else
              if pair is dissimilar then
                  tn $\leftarrow$ tn $+ 1$
              else
                  fn $\leftarrow$ fn $+ 1$
              end if
          end if
      end for
      precision $\leftarrow \frac{\text{tp}}{\text{tp}+\text{fp}}$
      recall $\leftarrow \frac{\text{tp}}{\text{tp}+\text{fn}}$
      f1_score $\leftarrow 2 \times \frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$
      if f1_score $>$ max_f_score then
          max_f_score $\leftarrow$ f1_score
          best_threshold $\leftarrow$ thr
      end if
  end for
  return best_threshold
```

[7] https://huggingface.co/sentence-transformers/all-roberta-large-v1

[8] https://huggingface.co/datasets/nyu-mll/glue/viewer/stsb/train

concept to identify paths leading to the other. This provides a total of 728 paths with various depths from the three resources. Human annotators then evaluate the relevance of the retrieved paths based on a binary value indicating if the path is relevant to the given AR or not. The cumulative F1-score at each depth is computed based on the total number of relevant paths retrieved up to that depth. The depth with the highest cumulative F-score is chosen as the optimal threshold. Accordingly, the threshold of $\alpha = 5$ yielded the highest score.

A.3.4 Filtering semantic relations.

A total of 7959 unique relation types are extracted. Please note that similar relation types like “leads to”, “leads” and “can lead to” are counted as different relation types, as we only consider surface-level counts. To exclude arbitrary paths between concepts only relation types with a frequency greater than *m=3* are considered. This yields a total of 1488 unique relation types. However, as can be seen in Table 7, manual analysis revealed similarities among certain tuples; for example, the relation type “influences” is similar to other relations like “contributes to”, “leads to”, and “results in”.

Some concepts are directly related through single relation type (one-hop path), while others are indirectly connected via paths involving multiple relation types (multi-hop). See examples in Table 6. The length of these paths ranges from 1 (indicating direct links between concepts) to 5 (the maximum search depth), with an average path length of 1.9.

A.3.5 Extracting keywords encoding semantic relation types from Wikipedia.

The AllenNLP semantic role labeling (SRL)[9] is used to parse sentences and assign semantic roles to each word. This enables to extract phrases linking the concepts of interest along the subject-predicate structure of the sentences. To mention, if one concept is identified as the *agent* and another as the *patient*, the phrase denoting the action performed by the agent on the patient is used as the relation type between them. Consider the concepts *exercise* and *cardiovascular diseases* in the sentence:

> *According to the American Heart Association, exercise reduces the risk of cardiovascular diseases, including heart attack and stroke.*

Below is the output of SRL for this sentence (the concepts are highlighted in light blue while the keywords representing the relation type are highlighted in red): {‘verbs’: [{‘verb’: ‘According’, ‘description’: ‘[V:According] to the American Heart Association , exercise reduces the risk of cardiovascular diseases , including heart attack and stroke’, ‘tags’: [‘B-V’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]}, {‘verb’: ‘reduces’, ‘description’: ‘[ARGM-ADV: According to the American Heart Association] , [ARG0: exercise] [V: reduces] [ARG1: the risk of cardiovascular diseases , including heart attack and stroke]’, ‘tags’: [‘B-ARGM-ADV’, ‘I-ARGM-ADV’, ‘I-ARGM-ADV’, ‘I-ARGM-ADV’, ‘I-ARGM-ADV’, ‘I-ARGM-ADV’, ‘O’, ‘B-ARG0’, ‘B-V’, ‘B-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’]}, {‘verb’: ‘including’, ‘description’: ‘According to the American Heart Association , exercise reduces the risk of [ARG2: cardiovascular diseases] , [V: including] [ARG1: heart attack and stroke]’, ‘tags’: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘B-ARG2’, ‘I-ARG2’, ‘O’, ‘B-V’, ‘B-ARG1’, ‘I-ARG1’, ‘I-ARG1’, ‘I-ARG1’]}], ‘words’: [‘According’, ‘to’, ‘the’, ‘American’, ‘Heart’, ‘Association’, ‘,’, ‘exercise’, ‘reduces’, ‘the’, ‘risk’, ‘of’, ‘cardiovascular’, ‘diseases’, ‘,’, ‘including’, ‘heart’, ‘attack’, ‘and’, ‘stroke’]}

We navigate through the SRL output to identify the predicate-argument structures connecting both concepts (*exercise* and *cardiovascular diseases* in this case). We then use predefined rules to extract keywords encoding the semantic relations existing between the concepts. To mention, if one concept is part of *ARG0* and the other being part of *ARG1*, the predicate term is used as the relation type. In the example output above, the predicate term representing the semantic relation type is *reduces*. More examples are provided below. The pair of concepts

[9] https://docs.allennlp.org/v0.9.0/api/allennlp.models.semantic_role_labeler.html

ADUs	Components, Chain of Hyperlinks
ADU1 - Trump tax cut is the biggest since Ronald Reagan; **ADU2** - It will create tremendous numbers of new jobs	**Chain of Hyperlinks for the components, Tax (C), Jobs (A):** - Job → Working hour system → Income tax → Tax - Job → Labor economics → Economic policy → Tax - Job → Unemployed → Tariff → Tax
ADU1 - Clinton is going to approve one of the biggest tax cut in history; **ADU2** - Clinton is going to drive business out	**Chain of Hyperlinks for the components, Business (C), Tax cut (A):** - Business → Adam Smith → Economic theory → Tax cut - Business → Adam Smith → Neoliberalism → Tax cut - Business → Corporate tax → Effect of taxes and subsidies on price → Tax cut

Table 5: Examples showing the connection between ADUs via the chain of hyperlinks linking their components.

are highlighted in light blue and the relation type highlighted in red:

1. **Concept Pair: Exercise, Cardiovascular Diseases**
   - **Semantic Relation:** increase
   - **Sentence:** "Low levels of physical exercise increase the risk of cardiovascular diseases mortality."
   - **Predicate structure:** [ARG0: Low levels of physical exercise] [V: increase] [ARG1: the risk of cardiovascular diseases mortality].

2. **Concept Pair: Exercise, Cardiovascular Profiles**
   - **Semantic Relation:** leads
   - **Sentence:** "Studies have shown that since heart disease is the leading cause of death in women, regular exercise in aging women leads to healthier cardiovascular profiles."
   - **Predicate structure:** Studies have shown that [ARGM-CAU: since heart disease is the leading cause of death in women], [ARG0: regular exercise in aging women] [V: leads] [ARG2: to healthier cardiovascular profiles].

3. **Concept Pair: Innovation, Economy**
   - **Semantic Relation:** is
   - **Sentence:** "Given the noticeable effects on efficiency, quality of life, and productive growth, innovation is a key factor in society and economy."
   - **Predicate structure:** [ARGM-ADV: Given the noticeable effects on efficiency , quality of life , and productive growth], [ARG1: innovation] [V: is] [ARG2: a key factor in society and economy]

4. **Concept Pair: Sustainable Energy, Renewable Energy**
   - **Semantic Relation:** involves
   - **Sentence:** "Sustainable energy involves increasing production of renewable energy, making safe energy universally available, and energy conservation."
   - **Predicate structure:** [ARG2: Sustainable energy] [V: involves] [ARG1: increasing production of renewable energy , making safe energy universally available, and energy conservation]

A.3.6 External Resource Evaluations

Ontology and Wikipedia: We analyse the three resources to showcase their contributions in terms of coverage and the quality of connections.

Coverage. The aim is to show the proportion of pairs connected exclusively by one resource but not by others. To this end, we randomly select 500 unconnected pairs from each resource and generate a heatmap illustrating the ratio of pairs exclusively connected by each resource compared to the others to identify which resource is most effective in covering concepts absent in others. On average, Wikipedia covers 37% of pairs unconnected in both WordNet and ConceptNet, while only 7% of the concepts missing in Wikipedia are covered by both WordNet and ConceptNet. Please note that pairs of concepts connected by relation types occurring less than three times are considered unconnected.

Connection Quality. We further analyse the quality of the paths by ranking component pairs based on the number of paths linking them from each respective resource. From this ranking, we select the top 25 most connected and 25 least connected pairs from each resource for detailed evaluation. Two annotators independently assess the relevance of these paths by assigning binary labels, reflecting their subjective evaluations of the paths' pertinence to the AR between the ADUs.

The evaluation shows that Wikipedia emerges as the top-rated source for both well-connected and least connected paths, achieving an F1 score of 0.73. ConceptNet follows closely with an F1 score of 0.71, while WordNet has an F1 score of 0.68, indicating its comparative effectiveness in providing relevant connections.

GPT-generated paths: As shown in Table 2, configuration utilising GPT-generated paths show higher accuracy but lower precision. Of the total errors observed, 79% are identified as false positives for approaches using GPT-generated paths in predicting AR, while the average false positive rate for the other three external resources is 53%. To further investigate, we randomly select 20 errors and engage two human annotators to jointly analyse the paths connecting the pair of ADUs.

Out of the 20 errors, the paths for the 14 of the errors are categorised as contextually irrelevant for the ADU pairs. The primary reason cited by the annotators for the irrelevant paths indicates that while the generated paths make logical sense and provide valid lines of reasoning between the ADUs, there were no AR between these ADUs as originally annotated in the dataset.

For example, consider the pair of ADUs *"Researches into humanities and art still need large amount of money"* and *"a government should spare effort on young children education as well as universities"*, taken from the argument graph depicted in Figure 2 (taken from AAEC dataset). GPT identified the following semantic relation paths linking the concepts *"money"* and *"young children education"*:

- **money** → *facilitates* → **technology adoption** → *enables* → **digital literacy programs** → *encourages* → **young children education**

- **money** → *stimulates* → **philanthropic endeavors** → *cultivates* → **community partnerships** → *fosters* → **early childhood learning opportunities**

- **money** → *fuels* → **economic growth** → *stimulates* → **job creation** → *expands access to* → **early childhood education**

- **money** → *drives* → **philanthropic activities** → *funds* → **charitable organisations** → *supports* → **early childhood education initiatives**

- **money** → *empowers* → **local communities** → *cultivates* → **community engagement** → *enhances* → **early childhood learning environments**

- **money** → *encourages* → **resource allocation** → *drives* → **research and development** → *inspires* → **pedagogical advancements**

Despite these two ADUs not being linked by AR in the gold dataset, the paths between the concepts they address mimic the paths typically associated with ADUs involving AR. However, the reasoning conveyed by these paths is categorised as unintended, as they involve reasoning diverging from the original argument, and the AR between these ADUs is absent in the gold dataset.

The same applies to the paths identified for the concepts *money* and *future* addressed by the pair of ADUs: *"Researches into humanities and art still need large amounts of money"* and *"both are crucial on the way to a brighter future"*.

- **money** → *allows for* → **travel experiences** → *impacts* → **cultural enrichment** → *shapes* → **future memories**

- **money** → *allows for* → **excessive spending** → *impacts* → **short-term pleasure** → *shapes* → **future goals**

- **money** → *initiates* → **investment opportunities** → *promotes* → **financial stability** → *contributes to* → **future security**

- **money** → *used for* → **educational funding** → *influences* → **career advancement** → *impacts* → **future**

- **money** → *is used for* → **investment in property** → *helps in* → **wealth accumulation** → *contributes to* → **future**

- **money** → *is used for* → **infrastructure development** → *helps in* → **urban planning** → *contributes to* → **future city growth**

- **money** → *allows for* → **business expansion** → *impacts* → **economic prosperity** → *shapes* → **future**

- **money** → *leads to* → **business expansion** → *is linked to* → **economic growth** → *impacts* → **future prosperity**

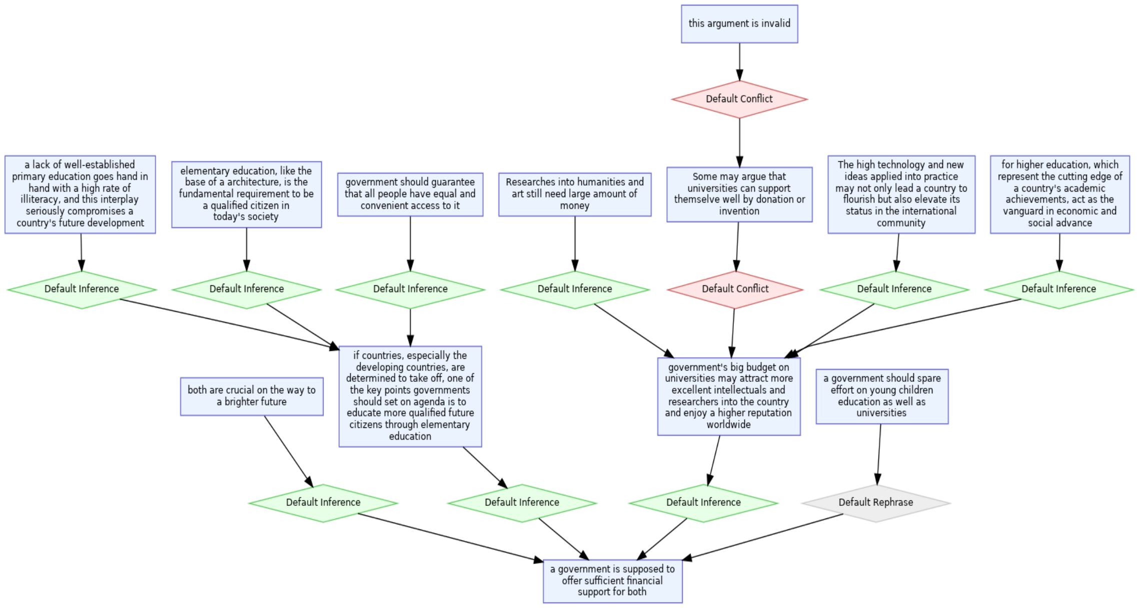
this argument is invalid
Default Conflict
a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay seriously compromises a country's future development
elementary education, like the base of a architecture, is the fundamental requirement to be a qualified citizen in today's society
government should guarantee that all people have equal and convenient access to it
Researches into humanities and art still need large amount of money
Some may argue that universities can support themselve well by donation or invention
The high technology and new ideas applied into practice may not only lead a country to flourish but also elevate its status in the international community
for higher education, which represent the cutting edge of a country's academic achievements, act as the vanguard in economic and social advance
Default Inference
Default Inference
Default Inference
Default Inference
Default Conflict
Default Inference
Default Inference
both are crucial on the way to a brighter future
if countries, especially the developing countries, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education
government's big budget on universities may attract more excellent intellectuals and researchers into the country and enjoy a higher reputation worldwide
a government should spare effort on young children education as well as universities
Default Inference
Default Inference
Default Inference
Default Rephrase
a government is supposed to offer sufficient financial support for both

Figure 2: Example argument graph.

- **money** → *is essential for* → **scientific research** → *contributes to* → **technological advancement** → *shapes* → **future innovation**

A.4 GPT for Path Generation and AR Prediction

A.4.1 Experimental Settings

We utilise the chat completion configuration of ChatGPT-4 for two tasks: (a) generating the chain of semantic relation between ADU components, and (b) predicting AR.

1. **Configurations:** We use GPT-4 based on `gpt-3.5-turbo-instruct`. We set a maximum token limit of 2048, a temperature of 0.7, a top-p probability of 0.9.

2. **Prompts Strategy:** We explored two strategies: zero-shot and few-shot prompts. In the zero-shot setting, only instruction based prompts without examples are used. We also try few-shot setup, where specific examples are provided as part of the instruction. Interestingly, our analysis revealed that the example-based experiment achieved a 1.3%, 2.1% higher score compared to the zero-shot prompt in the AR prediction and path generation, respectively. As a result, our experiment is based on example-based prompting. We create prompt templates that include instructions and two examples randomly selected from a list of examples. These examples consist of ADU pairs, concept pairs identified from the ADUs, and paths obtained from three external resources. The placeholder variables in the template are replaced with the ADUs, concepts, and paths.

Prompt Design for Path Generation. GPT-4 is tasked with generating paths between components of ADUs using the following template:

```
You are a model trained to
identify chains of semantic relations
between a pair of concepts derived
from two sentences (ADU1 and ADU2).
Given concepts c1 and c2 extracted from
ADU1 and ADU2 respectively,
your goal is to identify chains of
semantic relation types connecting
these concepts. These relations may
include meronymy, hypernymy,
hyponymy, cause-effect, or any other
valid semantic relation. Concepts are
often indirectly linked via
intermediate concepts and their
relations. Include both direct and
indirect paths between the concepts
whenever possible,  using only
the context provided by the ADU pairs.
Provide up to 10 paths if possible;
otherwise, return an empty list.
Each relation type should be
represented as a tuple in the format
(concept1, relation type, concept2).
```

```
For indirect paths involving
multiple tuples, return them
as a list of tuples.
Example 1:between the concepts
"USA" and "NAFTA" identified from
the pairof ADUs
"USA is in deep trouble"
and "NAFTA agreement is defective",
a valid list of paths could be,
[[("USA,part-of,NAFTA")"],
["(USA, has, trade deal),
(trade deal, instance of, NAFTA")].
Example 2: between the concepts
{c1} and {c2} identified from the
pair of ADUs {ADU1} and {ADU2},
the list of paths should include,
{list_path}.
Provide your answer as a python list.
```

Note: In Example 1, we show an actual example, but it should be a placeholder variable in the prompt template, as shown in Example 2.

Prompt Design for Zero-Shot AR Prediction: We prompt GPT-4 to classify the relationship between the ADUs as supporting, contradicting, or having no clear AR using the following prompt template.

```
You are a 3-class classifier model
tasked with assigning a label
to the argument relation between
two argument units
(argument 1 and argument 2).
Classify the following
pair of arguments,
argument 1: {ADU_1}
argument 2: {ADU_2},
into:
"support" (if argument 1 supports
argument 2),
"contradict" (if argument 1 attacks
argument 2),
and "None" (if no argument relation
exists between
argument 1 and argument 2).
Please enter:
1 - for support,
2 - for contradict,
0 - for None relation.
Examples from each argument
relation types are provided below:
Example 1: the argument relation between
the argument "people feel, when they have
been voicing opinions on
different matters, that they
have been not listened to",
and the argument "people
feel that they have been treated
disrespectfully on all sides of the
different arguments and disputes
going on" is support, and  hence
prediction label is 1.
Example 2: The argument relation
between "there would be no non-tariff
barriers with the deal  done with
the EU" and the argument
"there are lots of
non-tariff barriers
with the deal done with the EU"
is contradiction, and
hence prediction label is 2.
```

Note: We use the actual examples to show support and contradiction relations, which should be a placeholder variable in the final prompt template.

A.5 Multi-Network Architectures

The encoder blocks within the multi-networks are constructed using the HuggingFace implementation of BERT (bert-base-uncased) [10]. In all configurations, we utilise 8 attention heads, which is a common feature in standard transformer implementations. This design choice allows the model to attend to different parts of the input sequence simultaneously, enhancing its ability to understand and represent complex relationships within the data.

A.5.1 Attention Mechanisms in Multi-Network Architectures

The Triplet Network architecture is aimed to encode the individual components of ADUs as well as the external knowledge paths connecting them. The architecture consists of three sub-networks, each focusing on a different aspect of the input:

- **Sub-Network 1**: Encodes the premise.
- **Sub-Network 2**: Encodes the conclusion.
- **Sub-Network 3**: Encodes the paths between components of ADUs.

Two attention layers are used to attend to the alignment between the inputs (premise, conclusion and external knowledge).

[10] https://huggingface.co/google-bert/bert-base-uncased

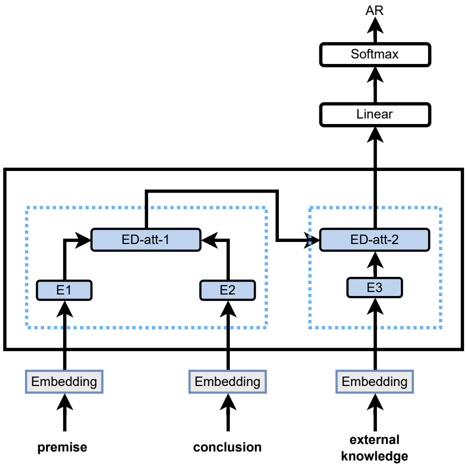

Figure 3: Triplet-networked with attention layers.

1. **First Attention Layer**: This layer attends to the alignment between the premise and conclusion based on the outputs of Sub-Networks 1 and 2, respectively.
2. **Second Attention Layer**: Building upon the output of the first attention layer, this layer aligns the information from the first attention layer with the external knowledge provided by Sub-Network 3.

Finally, the outputs of the attention layers are passed through feedforward and a final linear classification layer to predict AR. We experiment with two configurations for representing the ADUs and the external knowledge as input to the attention layer: (a) using the final output of the [CLS] token and (b) using the mean of the last hidden layer of all tokens from BERT's output. Consistently, the mean of the last hidden layer of all tokens yields superior performance compared to the [CLS] token. Since our goal is a classification task that predicts the AR across three classes, we utilise cross-entropy loss based on the output of the linear classification layer. Triplet loss is not suitable for our specific task in either the Siamese or Triplet configurations.

A.5.2 Training Complexity and Additional Parameters

The inclusion of cross-attention layers and feedforward layers in both the Siamese and Triplet Network architectures introduces additional parameters and computational complexity over the baseline BERT model, which consists of 110 million parameters.

In the Siamese Network, cross-attention is applied between the premise-conclusion pair and external knowledge. The cross-attention mechanism contributes approximately **1.77 million parameters** (specifically, **1,769,424** parameters), while the subsequent feedforward layer adds around **2.37 million parameters** (specifically, **2,367,486** parameters). Additionally, the classification layer, which accounts for 3 output classes, adds about **2,304 parameters**. This brings the total parameter count for the Siamese model to approximately **116.5 million**, introducing about **4.14 million parameters** beyond the baseline.

The Triplet Network further increases complexity, employing two cross-attention layers between the premise, conclusion, and external knowledge, which contribute around **3.54 million parameters** (specifically, **3,538,944** parameters). The feedforward layers contribute approximately **4.74 million parameters** (specifically, **4,734,972** parameters), and the classification layer again adds **2,304 parameters**. This results in a total of approximately **123 million parameters**, which is around **8.28 million more** than the baseline.

The added cross-attention layers introduce additional computational steps by computing attention scores between inputs (premise, conclusion, and external knowledge), thereby increasing training complexity relative to the baseline model, which solely fine-tunes the BERT layers without external knowledge integration.

Path	Path	Path	Path
related to → leads to → affects	related to → related to	synonym	involves
related to	affects → associated with → impacts	synonym → related to	causes
has	is related to	leads to	is a → related to → related to
impacts	related to → involves	contains	is a → involves
is a → is a → is a	part of	related to → part of	causes → related to
influences → affects	antonym → hyponym → hyponym	associated with	involves → related to
leads to → results in	entails → entails	is a → has	can lead to
implies	related to → related to → related to	affects → influences	causes → leads to
has → includes	part of → includes	related to → includes	related to → related to → causes
supports	synonym → hypernym → hyponym	entails → involves	related to → entails
causes → affects	is a → belongs	is associated with → involves	regulate
related to → impacts	can result in	is an umbrella term → is related to	leads to → involves
found in	administers → impacts → involves	affiliated with → associated with	aids → helps → helps
developed through → facilitated by → leads to	discussed in → is a → lead to	are → show → can lead to → can result in	assessment of → measure of → related to
associated with → is a type of → can be	be used for → have quality	can be obtained → is extended for → is a type of	occur in → experiencing → necessitate
can provide → may lead to → changes	common in → generally involves	includes → example of	determines → affects
empowerment through → instance of	entails → sustain	entails → includes → involves	entails → is a → is a → is a
entails → is required for → can lead to → can result in	experienced → includes	often favours → which stems from	fosters → crucial for
give → way to	impacts → evaluates	influences → lead to → affect	influences → are reflected by
influences → is achieved by	influences → importance of → includes	involves → involvement of → can come under	involves → is represented by
involves → brings → used for	is a factor in → generates → can include	symptom of → includes → has code	is a type of → may require → is associated with
is a → is delivered through → facilitates	is essential to → has an impact → results in	is important for → used in → opportunity for	is involved in → has phase → is type for
is often accompanied by → is similar to	is often associated with → has effects on → are linked to	related → shapes → contribute to → are crucial for	required → necessary for
supported by → promotes → reduces → are important	is the goal of → can include	is type of → can involve → is related to	is a → involves → relieves
live in → has	may bring → followed by → result in	may lead to → requires → found	necessitates → involves → category of
offer → facilitate → contributes to → aids in	opposite of → causes → leads to	organised by → hold	participates in → can involve
provide → attend	provides challenge in	provides → enable	provides → includes → develops → can lead to
refers to → impacts → affects	related to → can lead to → results in → results in	related to → improves → essential → crucial for	related to → indicates → compared to
leads to → is likely to → often achieved by	represents → causes	require → achieved by → help	shares → comprises of
duration → has value	used for → associated with → part of	convey → interpreted by → part of → makes up	requires → causes → leads → necessary

Table 6: Examples of semantic relation paths.

Relation	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
related to	involves	hyponym	antonym	synonym	is a
has	results in	affects	related term	leads to	can lead to
causes	entails	associated with	part of	includes	hypernym
influences	impacts	is related to	contributes to	include	is a type of
instance of	can result in	requires	related	connected to	contains
have	require	can be	involve	used for	implies
consists of	versus	are	lead to	greater than	affect
influence	entailment	type of	causes desire	linked to	cause effect
opposite of	relates to	is essential for	is similar to	impact	may lead to
supports	provides	can involve	is crucial for	result in	is part of
cause	essential for	may result in	symptom of	is a form of	comparison
facilitates	enhances	motivated by	contribute to	can cause	similar
used in	experience	is important for	enables	influenced by	drives
at location	provide	are part of	percentage	may involve	comprises
synonym of	opposite	indicates	describes	attribute	attend
refers to	is	can include	determines	promotes	has instance
use	participate in	entails action	treated with	utilises	measured by
shapes	pertains to	is connected to	necessitates	encourages	improves
antonym of	is used in	similar to	measures	is used for	represents
chain map	offers	is influenced by	treat	may cause	of
negation	has context	shape	consist of	has property	example of
motivates	are associated with	equals	can affect	location	has quality
enhance	relate to	affected by	undergo	may include	contributes to
belongs to	can influence	found in	addresses	impact on	create
seek	possess	increases	can impact	receive	compares
opposes	member of	feature	subset of	concerns	is required for
derived from	is a part of	has attribute	resulted in	comprise	is equivalent to
treatment for	used by	activity	treatment	regulates	correlates with
enable	produces	is necessary for	triggers	target	ensures
inspires	correlated with	impacted by	inspire	helps in	has duration
need	is a factor in	component of	is known for	modifies	related term of
measure	treated by	is less than	characteristic of	has numeric value	covers
is about	is a symptom of	employs	entail	located in	has part
located near	shows	are crucial for	focuses on	engage in	depends on
pursue	is needed for	brings about	motivated by goal	cause of	can have
can	associated with	are related to	range	involved in	utilise
targets	means	attribute of	benefits	characterised by	measured by
spouse of	is a measure of	side effect of	comparative of	can be influenced by	is vital for
has member	occurs in	evaluate	implement	allows for	has symptom
is equal to	less than	belong to	linked to	involves	encourage
fosters	component of	known for	capable of	is key to	helps
interact with	drive	constitute	relies on	comprises of	meronym
defines	generates	correlate with	determine	has subevent	represent
is an umbrella term	compare	has prerequisite	facilitate	desires	percentage of
a type of	acquired through	address	addressed	advocates for	agent
agent of	are important for	aligns with	aid in	are used in	assesses
attract	belongs to group	belong to	boosts	achieved through	be found in
can be represented	can contribute to	can create	can enhance	can develop into	can lead to
can occur in	can require	can stimulate	capability of	category of	caused by
combined with	complication of	concept in	connects	conceptually related to	deals with
essential for	establish	evaluated by	examines	example of	exhibit
has percent	has range	has impact on	have activity level	helps to get more	helps gain
holonym	impacts result in	implemented by	imposes	indicate	induces
inhibits	is a medication	is a metric for	is a side effect of	is a source of	is a subclass of
is a symptom of	is a unit of time	is a way to	is beneficial for	is critical for	is defined by
is fundamental for	is funded by	is greater than	is key to	is opposite of	is perceived as
is quantified by	is significant for	has value	is the target of	is treated with	is used to assess
is treated by	lack of	lacks	live in	location of	made by some
made of	manifests as	negatively impacts	numerical value	often involve	outcome of
percentage value	play a role in	plays a role in	politician	possesses	prevents
process of	produce	promoted by	provide access to	provided by	qualifies
quantity	reduces	reflect	reflects	regulated by	rely on
restrict	results in state	show	stimulates	studies	suggests
superlative	to be gained by	tool for	treats	treatment includes	treated by
treatment involves	treatment with	trigger	utilised for	increased expression of	yield

Table 7: Examples of semantic relation types. We normalised (lower cased and expanded relation types like IsA, RelatedTo, HasProperty) the relation types for consistency across the external resources.