Chapter 6

External knowledge integration

This chapter explores feature-based and transfer-based approaches to integrate different kinds of external knowledge into argument mining models, and to investigates the impact of such integration on the performance of these models. The evaluation of these approaches and knowledge sources is conducted on both formal and informal argument datasets, including the Persuasive Essays dataset and the Quora dataset.

The research conducted in this chapter addresses the need for model-agnostic knowledge integration in end-to-end argument mining. It focuses on approaches that are not specific to a particular model architecture or subtask, but instead aim to provide general techniques that can be applied to various end-to-end argument mining models. As a result, model-specific approaches (*e.g.*, fine-tuning the BERT encoder with extra text) and approaches that target on a specific subtask (*e.g.*, utilising pre-trained BERT to boost relation identification) are not considered in this study. This broader exploration allows for a comprehensive understanding of the benefits and challenges associated with integrating external knowledge, providing insights into how it can enhance the performance and capabilities of argument mining models.

6.1 Approaches to external knowledge integration

External knowledge can be integrated into argument mining models in different ways, including as features to be added to the input text, and as pre-trained model weights through transfer learning, allowing models to benefit from additional information and context. Also, the knowledge to be integrated can be obtained from various sources. This section describes the feature-based and transfer-based approaches used in this thesis to incorporate different kinds of external knowledge.

6.1.1 Feature-based approach

In the feature-based approach, external knowledge is used as additional features that augment the input text, providing additional contextual or semantic information that can aid in argument mining. In the context of end-to-end neural models, these features are usually vector representations extracted from external knowledge sources, which are then combined with the input original representation of the input text. In this study, various sources of external knowledge are investigated, including syntactic information, discourse information, information in a curated knowledge graph, and information obtained from a pre-trained large language model.

Particularly, in this study, the investigation of pre-trained word embeddings as an additional feature for knowledge integration is not conducted. This is because subword level WordPiece embeddings Wu et al. (2016) are inherently incorporated by the pre-trained BERT encoder in the proposed biaffine and GNN-based parsers. This decision is justified by the fact that pre-trained word embeddings have already been widely acknowledged as a powerful external knowledge source, and that WordPiece embeddings have been proved to be effective in many NLP tasks Gururangan et al. (2020), Raffel et al. (2020). As a result, pre-trained word embeddings are considered as a a standard and integral component of the input representation to argument mining models, instead of a variable to be examined.

6.1.1.1 Syntactic information

Syntactic information plays a crucial role in analysing the internal structure of sentences, which can be beneficial for determining segment boundaries in argument mining. Moreover, it is expected to assist in resolving ellipsis and anaphora (Dalrymple et al., 1991, Lappin and Leass, 1994), which are common phenomena in argumentative texts, particularly informal arguments. Ellipsis refers to the omission of certain words or phrases that can be inferred from the context, while anaphora involves the reference of a word or phrase to a previously mentioned antecedent. Therefore, syntactic parses of the target text is investigated as a source of external knowledge for argument mining.

To incorporate syntactic information into the argument mining models, the syntactic parse tree of each sentence in the target text is obtained using the Stanford CoreNLP

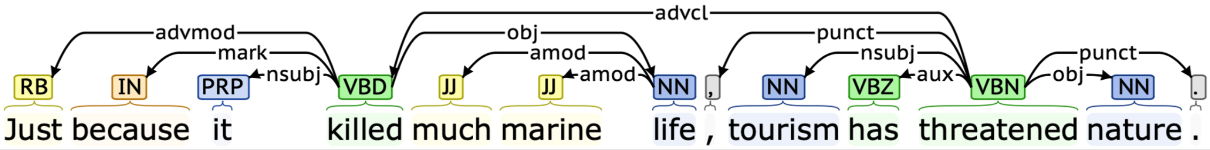

Figure 6.1: An example syntactic parse tree produced by Stanford CoreNLP (Manning et al., 2014).

syntactic dependency parser (Manning et al., 2014). The syntactic parse tree provides a hierarchical representation of the sentence structure, illustrating the syntactic relationships between tokens. An example of such a tree is illustrated in Figure 6.1.

For each token, its path to the root node in the parse tree is encoded as a list of tuples in the form of ($edge\ label, node\ index$) along that path. For example, in Figure 6.1, the token "life" has the path $[(obj, 4), (advcl, 11), (root, 0)]$, with one-hot encoding applied to the edge labels.

6.1.1.2 Discourse information

While syntactic information reveals the structure of a sentence through token-level dependencies, discourse information describes the overall structure of an entire document through segment-level dependencies. Discourse structures can be represented using the RST Rhetorical Structure Theory (RST) framework, which describes how different seg-

[Farmington police had to help control traffic recently]$_1$ [when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.]$_2$ [The hotel's help-wanted announcement - for 300 openings - was a rare opportunity for many unemployed.]$_3$ [The people waiting in line carried a message, as refutation, of claims that the jobless could be employed if only they showed enough moxie.]$_4$ [Every rule has exceptions,]$_5$ [but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,]$_6$ [not laziness.]$_7$

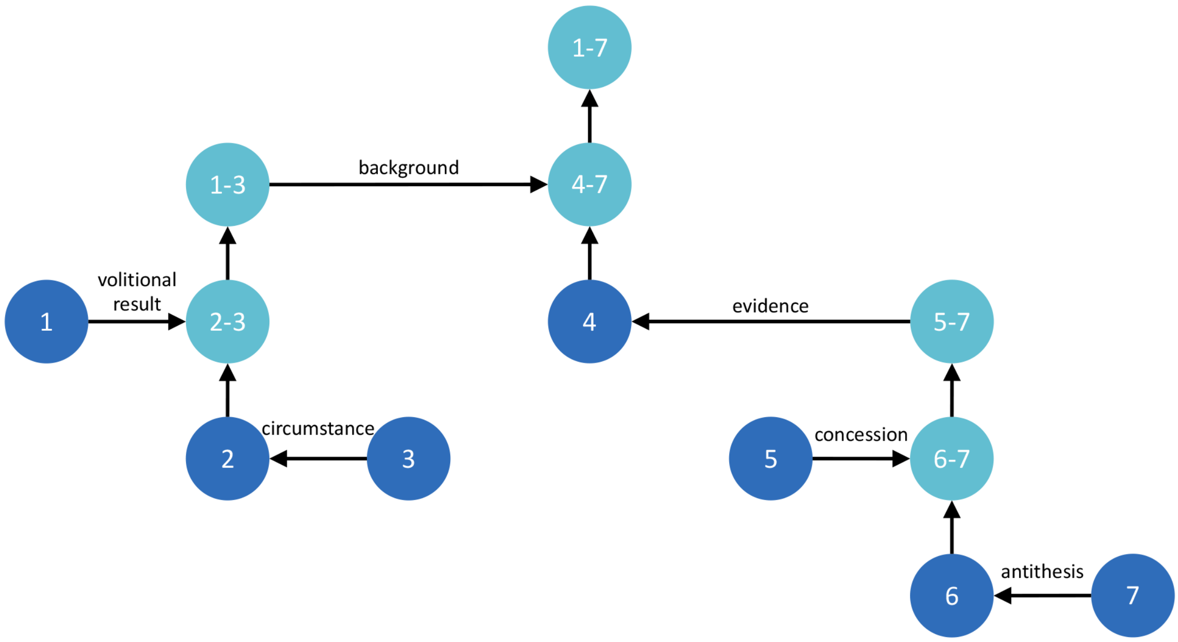

Figure 6.2: An example RST parse tree. Numbers are indices of text segments.

ments in a document are hierarchically organised into larger discourse units. The relations between these units, known as rhetorical relations, capture the discourse relationships and coherence within the text. An example of an RST tree is illustrated in Figure 6.2.

RST trees have been acknowledged to be useful to many NLP tasks (Marcu, 2000, Burstein et al., 2001), and are considered to hold strong resemblance to argument structures (Accuosto and Saggion, 2019). Some relations in the RST framework are indicative of relation categories in the annotation scheme proposed in this thesis. For instance, in Figure 6.2, the EVIDENCE relation between segments [5-7] and segment 4 may hint a SUPPORT relation between segment segment 6 and segment 4, and the CONCESSION relation between segment 5 and segments [6-7] can be a signal of segment 5 undercutting the relation between segment 6 and segment 4. Therefore, discourse parses obtained from an RST parser are investigated as a source of external knowledge to enhance argument mining.

The RST tree of each document in the target dataset is obtained using the RST parser proposed in Wang et al. (2017). Since the minimal unit in an RST tree is segment, each token in a segment shares the same discourse information. To encode the information in an RST tree, the path of each segment to the root node is represented as a list of tuples in the form of $(edge\ label, target\ unit\ span)$ along that path. Particularly, the unlabelled relation between a nucleus (*e.g.*, segments [6-7]) and its head (*e.g.*, segments [5-7]) in an RST tree takes the label of "span" in the encoding process. For example, segment 5 has the path $[(concession, 78-107), (span, 73-107), (evidence, 46-72), (span, 46-107), (span, 1-107), (root, 0)]$, with one-hot encoding applied to the edge labels.

6.1.1.3 Knowledge graph

Knowledge graphs typically offer structured information about entities, concepts, and events, as well as the relationships between them. ConceptNet Speer et al. (2017) is a large-scale, open-source curated knowledge graph that represents concepts as nodes and their relationships as edges. These relationships include both bi-directional (*i.e.*, symmetrical) relations, such as "located near", "related to", and "synonym", and uni-directional (*i.e.*, asymmetrical) relations, such as "is a", "part of", and "used for". An example of a partial graph of the term "argument" in ConceptNet is illustrated in Figure 6.3.

The common-sense knowledge in ConceptNet is expected to augment argument mining by supplying semantic information about concepts that may not be explicitly present in the training data, as well as by bridging the gap between concepts that are not directly stated but are understood within arguments. For example, consider the argument, "Carlos does ten minutes of jumping rope every day and his cardiovascular health has significantly improved." Jumping rope may appear rarely in the training data, and the connection between jumping rope and cardiovascular health might not be apparent from the training

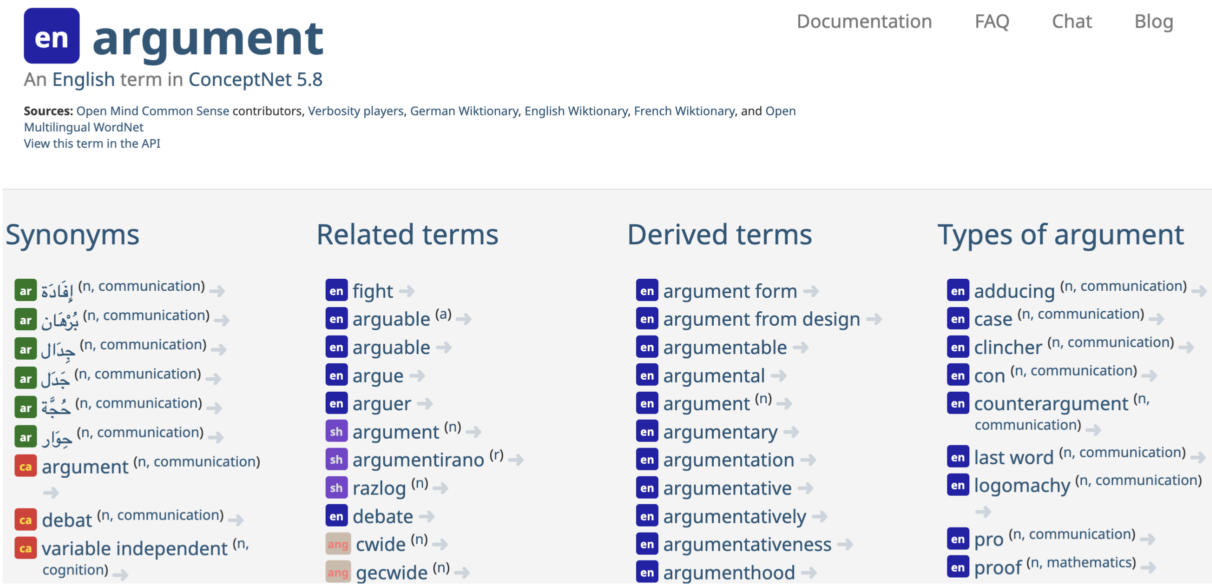

Figure 6.3: A partial graph of the term "argument" in ConceptNet. Four kinds of relations are displayed, namely "synonym", "related to", "derived from" (incoming edge), and "is a" (incoming edge).

data alone. However, ConceptNet explicitly states that jumping rope is "a type of" good exercise "used for" cardiovascular health. Such common-sense knowledge could be invaluable to argument mining models, and as a result, ConceptNet is explored as an external knowledge source in this study.

ConceptNet represents concepts as words and phrases in various languages. Given that the target text in this study is in English, and that the proposed AMDP approach operates at the token-level, only single-word English concepts are included. Furthermore, only the undisambiguated form of words from ConceptNet are used, as the proposed AMDP approach does not involve word sense disambiguation.

Each token's ConceptNet information is encoded as a list of tuples in the form $(relation\ label, linked\ concept, edge\ direction)$. For instance, the ConceptNet information for "argument" from Figure 6.3 would be encoded as $[(debate, synonym, bi-directional), (fight, related\ to, bi-directional), (argumentable, derived\ from, incoming)$, $(adducing, is\ a, incomining), \ldots]$. These concepts and relation labels are then subject to one-hot encoding.

6.1.1.4 Pre-trained large language model

Pre-trained large language models, exemplified by GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023), have exhibited exceptional capabilities in understanding and generating human-like text. Developed on a broad range of internet text with a Transformer-based architecture, these models have made significant strides in numerous natural language processing tasks including question answering, text completion, and text generation.

In the context of argument mining, pre-trained large language models are expected to serve as powerful external knowledge sources. The training of these models on massive amounts of data enables them to implicitly acquire a wealth of world knowledge, diverse linguistic patterns, and robust reasoning capabilities. However, GPT-3.5, which is powerful and freely available, operates as a generative language model and exclusively outputs natural language text. Consequently, in order to harness GPT-3.5 as an external knowledge source for argument mining, a structured approach is needed. Specifically, GPT-3.5 can be employed to generate responses to specifically designed prompts that simulate argument mining tasks. Subsequently, these generated responses can be processed to extract information pertinent to the task at hand.

The structured approach employed in this study to utilise GPT-3.5 for argument mining is inspired by the feedback received from the annotators of the Quora dataset. During the annotation process, they find it beneficial to initially identify the main statements in an argument, which convey the author's primary stance. Accordingly, the entire text of an argument, coupled with the question "Which exact segments in the text above are the main statements of this argument?" is presented as the prompt to GPT-3.5.

For every segment that GPT-3.5 returns, the complete argument along with questions including "Which exact segments in the text above support the claim that '[segment]'?" and "Which exact segments in the text above attack the claim that '[segment]'?" are subsequently presented as new prompts to GPT-3.5. There are instances when a single segment is identified by GPT-3.5 as both a supporting and attacking segment for the same target segment, which is a structure typically not permissible in most annotation schemes. In such cases, the prompt "Does the claim '[overlapping segment]' support or attack the conclusion '[target segment]'?" is posed to GPT-3.5, and the relationship between the overlapping segment and the target segment is determined based on GPT-3.5's response. This process of support/attack prompting is iteratively conducted up to three times, or until no new segments are returned.

Finally, an argument graph is constructed using this structured approach, which is detailed in Algorithm 1.

Specifically, due to the inherent limitations of GPT-3.5, the segments returned by the model sometimes do not exactly match those present in the original argument. Thus, for each returned segment, an initial step is taken to check if it exactly matches a segment in the original argument. If not, GPT-3.5 is fed with a prompt comprising the original argument and the instruction "Please find the exact segment in the original text for '[segment]'." This step is repeated until an exact match is found. The initially returned segment is then replaced by this exact match for subsequent steps in Algorithm 1.

In an effort to fully exploit the few-shot learning abilities of GPT-3.5, ten documents randomly selected from the training set of each target dataset are input to GPT-3.5 using the prompting strategy prior to the generation of the argument graph. Additionally, to

Algorithm 1 Argument Mining with GPT-3.5

```
Require: Argument A
Initialize ArgumentGraph
Prompt_main ← "Which exact segments in the text above are the main statements
    of this argument?"
Segments ← GPT(A + Prompt_main)
Add Segments to ArgumentGraph
CurrentSegments ← Segments
for i = 1 to 3 do
    Initialize DepSegments
    for each CurrentSegment in CurrentSegments do
        Prompt_sup ← "Which exact segments in the text above support the claim
            that '[CurrentSegment]'?"
        SupportingSegments ← GPT(A + Prompt_sup)
        for each SupportingSegment in SupportingSegments do
            if SupportingSegment in ArgumentGraph then
                Remove SupportingSegment from SupportingSegments
            end if
        end for
        Prompt_att ← "Which exact segments in the text above attack the claim
            that '[CurrentSegment]'?"
        AttackingSegments ← GPT(A + Prompt_att)
        for each AttackingSegment in AttackingSegments do
            if AttackingSegment in ArgumentGraph then
                Remove AttackingSegment from AttackingSegments
            end if
            if AttackingSegment in SupportingSegments then
                Prompt_op ← "Does the claim '[AttackingSegment]' support or attack
                    the conclusion that '[CurrentSegment]'?"
                OverlapCheck ← GPT(Prompt_op)
                if 'support' in OverlapCheck then
                    Remove AttackingSegment from AttackingSegments
                else
                    Remove AttackingSegment from SupprotingSegments
                end if
            end if
        end for
        Add SupportingSegments to ArgumentGraph with target segment index and
            relation label
        Add AttackingSegments to ArgumentGraph with target segment index and
            relation label
        Add SupportingSegments to DepSegments
        Add AttackingSegments to DepSegments
    end for
    CurrentSegments ← DepSegments
end for
return ArgumentGraph
```

mitigate potential bias that could be introduced by the order in which argument graphs are generated, each document in the target datasets is processed in an isolated session with GPT-3.5.

The information in the generated argument graph for each document in the target text is encoded in a manner similar to that of the discourse information, as described in Section 6.1.1.2.

6.1.2 Transfer-based approach

The transfer-based approach involves integrating external knowledge from datasets in other domains or tasks that are relevant to the target task. In the context of end-to-end neural models, this external knowledge is usually integrated as pre-trained model weights. The process starts with training models on a source task or domain, allowing for the accumulation of substantial external knowledge. The learned model weights from this initial training phase are then utilised as the starting point for training the model on the target task or domain. This study explores both cross-domain and cross-task transfer techniques.

Cross-domain transfer refers to the practice of applying knowledge learned from one domain to another. In the context of argument mining, different domains can refer to different genres of text, such as online forum debates, legal documents, and scientific literature. Each of these domains may have its own stylistic and structural nuances, which can potentially contribute valuable insights to argument mining. However, there is a necessity to accommodate potential variances in argument structures and language use across different domains, and annotation schemes utilised in various datasets. Such discrepancies could pose as potential obstacles to the transferability of knowledge, underscoring the importance of prudent dataset selection for successful cross-domain transfer.

Nonetheless, the options for argument mining datasets are limited, as discussed in Section 3.1. Thus, bearing in mind the relatedness to the Persuasive essays dataset and the Quora dataset in terms of annotation scheme, domain of source text, and dataset size, the Gold standard Toulmin dataset introduced in Section 3.1.2 is selected as the source dataset for cross-domain transfer in this study.

Cross-task transfer, on the other hand, involves applying knowledge learned from one task to another. The tasks in consideration should bear some relation to each other, and crucially, they should be solvable using the same neural model. As a result, syntactic parsing and discourse parsing are chosen as the source tasks for cross-task transfer in this study.

6.2 Experiment

An experiment is conducted to evaluate both the featured-based and the transfer-based approaches described above.

6.2.1 Dataset and experimental setup

In this experiment, evaluation is carried out using the Persuasive Essays dataset and the Quora dataset. The undercut-inclusive dependency representation proposed in Section 5.1 is employed, in conjunction with the GNN-based dependency parser introduced in Section 5.2.2.

For the feature-based approach, features described in Sections 6.1.1.1, 6.1.1.2, 6.1.1.3, and 6.1.1.4 are projected by an LSTM layer, and then are combined with the BERT-encoded representation through point-wise addition, respectively.

For the transfer-based approach, cross-domain transfer uses the Gold standard Toulmin dataset as the source domain dataset for pre-training the GNN-based parser. In the Gold standard Toulmin, argument relations are implicit but can be inferred according to the annotation scheme as follows:

- Backings are not related to any other components.
- The relation between a Premise and a Claim is always a Support.
- The relation between a Rebuttal and a Claim is always an Attack.
- The relation between a Refutation and a Rebuttal is always an Attack.

In the case of cross-task transfer, the data used to pre-train the GNN-based parser for the syntactic parsing task is generated by feeding the target datasets (*i.e.*, the Persuasive essays dataset and the Quora dataset) into the Stanford CoreNLP syntactic dependency parser. The pre-training data for the discourse parsing task is generated by feeding the target datasets into the RST parser by Wang et al. (2017), respectively.

The fine-tuning strategy, as proposed by Howard and Ruder (2018), is adapted for fine-tuning the GNN-based parser on the target datasets and task. In particular, the BERT encoder in the GNN-based parser is kept static during fine-tuning.

All other experimental setups adhere to the guidelines outlined in the experiments described in Sections 5.3 and 5.4.

6.2.2 Results

Results of the experiment on external knowledge integration are presented in Table 6.1. Based on the results, several major findings can be highlighted:

	Feature-based										Transfer-based					
	Base		Syn		Dis		CN		GPT		GST		SynPar		DisPar	
	C	R	C	R	C	R	C	R	C	R	C	R	C	R	C	R
PE	73.8	49.4	74.4	49.8	75.0	50.1	74.2	49.6	75.3	50.5	73.3	48.9	74.6	49.7	74.9	50.2
QR	66.2	45.8	67.8	47.0	68.0	47.5	67.0	46.4	69.2	48.5	66.7	45.6	67.4	46.6	68.6	47.9

Table 6.1: F_1 scores for approaches with different external knowledge integration methods. C = component identification, R = relation identification. Datasets: PE = the Persuasive essays dataset, QR = the Quora dataset. Feature-based approach: Syn = syntactic information, Dis = discourse information, CN = ConceptNet, GPT = GPT-3.5. Transfer-based approach: GST = cross-domain transfer with the Gold standard Toulmin dataset, SynPar = cross-task transfer with the syntactic parsing task, DisPar = cross-task transfer with the discourse parsing task. Base = GNN-based parser + undercut-inclusive representation.

- **Incorporating knowledge from GPT-3.5 outperforms all other methods**. Specifically, it achieves an F_1 score of 75.3% for component identification and 50.5% for relation identification on the Persuasive essays dataset. On the Quora dataset, GPT reaches an F_1 score of 69.2% for components and 48.5% for relations. This is indicative of the power of pre-trained large language models like GPT-3.5 in capturing rich semantic and contextual information in both formal and informal text, thus enhancing the performance of the model.

- In terms of the effectiveness of various external knowledge sources in the feature-based approach, there appears to be **a clear order of proficiency: GPT-3.5, followed by discourse information, syntactic information, and finally, ConceptNet**. This pattern underscores the idea that the efficacy of feature-based methods relies on the richness and applicability of the external knowledge they leverage, with GPT-3.5 managing to capture the most valuable information. Starting from the top, GPT-3.5 has demonstrated impressive results in numerous language understanding tasks. This language model has been trained on a diverse range of internet text, which allows it to capture not only syntactic and semantic patterns, but also discourse-level and pragmatic aspects of language use. Its capacity to understand context, nuance, and to draw upon a vast array of world knowledge makes it a formidable source of external knowledge for argument mining. Moving to discourse information, this form of external knowledge can offer valuable insights into how arguments are structured at the sentence level and beyond. By understanding the relationships between different parts of a text, the model can more effectively identify components and relations. Syntactic information, on the other hand, focuses on the grammatical structure within sentences. This information can be useful for understanding the roles and relationships of words and phrases within a sentence. However, as it operates on a lower level, its direct contribution to understanding the larger structures within arguments might be less compared to discourse information, hence the lower performance. Lastly, ConceptNet provides a knowledge graph of general world knowledge, which can be useful for enriching the model's understanding of semantic concepts and their interrelationships. Despite this, its performance is the lowest among the methods in the feature-based approach. This may be due to the fact that the type of knowledge it provides, while broad and diverse, may not be as directly applicable to the task of argument mining, which requires a deeper understanding of the structure and semantics of arguments.

- Among the methods in the transfer-based approach, **discourse parsing outperforms syntactic parsing as the source task, and the performance of using the Gold standard Toulmin dataset as the source dataset is close to the base performance, albeit slightly worse**. This is evident from the F_1 scores where DisPar has higher performance on both datasets, with a more pronounced improvement in relation identification compared to component identification. Conversely, SynPar's improvement is marginally higher in component identification than in relation identification. This suggests that cross-task transfer with the discourse parsing task provides more advantageous knowledge for the model, especially for relation identification. This finding is in line with the previous finding that discourse information is more useful than syntactic information as an additional feature, which suggests that knowledge which focus on understanding the high-level, holistic structure of texts can be particularly effective for argument mining. . The unsatisfactory performance of using the Gold standard Toulmin dataset for cross-domain transfer can be attributed to the difference in annotation schemes between the Gold standard Toulmin dataset and the two target datasets. The disparity in annotation schemes may lead to a misalignment in knowledge transfer, thereby impacting the model's performance.

- **The effectiveness of syntactic and discourse information varies between feature-based and transfer-based approaches**. For syntactic information, The Syn feature-based method outperforms the SynPar transfer-based method, indicating that direct feature integration can effectively harness syntactic structures for the task. Conversely, for discourse information, the DisPar transfer-based method surpasses the Dis feature-based method, suggesting that pre-training on a discourse parsing task provides more abstract and complex representations of discourse knowledge beneficial for argument mining. The difference in the effectiveness of syntactic and discourse information between feature-based and transfer-based approaches could be attributed to the intrinsic nature of the information they encode. Syntactic information, which is essential to understanding sentence-level structures, is highly specific and concrete.

When used as features in a feature-based approach, it can directly provide fine-grained structural cues to the model. This can be particularly useful for argument mining where identifying components often relies on understanding the specific grammatical structure of a sentence. In contrast, when syntactic parsing is used as a source task in a transfer-based approach, the pre-training process may not yield additional significant benefits. This is because the benefits of syntactic information lie in its concrete, immediate relation to the structure of sentences, which is already well captured through direct feature integration. Discourse information, on the other hand, provides an understanding of the relations and logical flow among larger text units. This type of information is more abstract and higher-level compared to syntactic information. As such, it may not be effectively harnessed through direct feature integration in the feature-based approach. However, when discourse parsing is used as a source task in the transfer-based approach, it allows the model to learn a more generalised understanding of discourse structures during the pre-training phase. This abstract and complex representation of discourse knowledge can be beneficial for argument mining, especially for the identification of relations that span across sentences or paragraphs.

- **External knowledge proves to be more beneficial for the Quora dataset than for the Persuasive essays dataset**. This is observable from the higher percentage increase in F_1 scores on the Quora dataset for the same methods as compared to the Persuasive essays dataset. This is very likely due to the fact that the more informal nature of the Quora dataset necessitates a greater extent of external knowledge to accurately parse the data. The persuasive essays dataset, which is presumably more formal and structured, likely adheres more strictly to standard grammatical conventions and argumentative structures. Because of this, models can often make accurate predictions based on structural patterns in the text and the use of particular lexical cues associated with argumentative components and relations. In contrast, the Quora dataset is likely to be more informal and variable in style and structure. The argument structures may be less explicit and more intertwined with colloquial language and individual writing styles. In such cases, external knowledge becomes more crucial in accurately interpreting the text and identifying components and relations. The use of external knowledge can help the model better understand the semantics and pragmatics of informal and potentially ambiguous language expressions, and it can provide additional contextual information that is not immediately available in the text itself. For instance, discourse features can aid in deciphering the implicit argument relations in less structured text, while the knowledge from GPT-3.5 can help comprehend the underlying meaning and context of colloquial or unconventional expressions in diverse topics.

In conclusion, most of the methods investigated in this study to integrate external knowledge for argument mining boost the model's performance. The key to success appears to lie in the alignment of the knowledge source and the integration technique with the nature of the task and the data at hand. Therefore, it is crucial to carefully choose the type of knowledge and the integration method to maximise the performance gains.

6.2.3 Summary

This experiment provides an in-depth evaluation of feature-based and transfer-based approaches for the integration of external knowledge into argument mining on the Persuasive essays and Quora datasets. For the feature-based approach, various sources of external knowledge are investigated as additional features, including syntactic information, discourse information, information obtained from ConcetNet, and information obtained from GPT-3.5. For the transfer-based approach, the Gold standard Toulmin dataset is used as the source dataset for cross-domain transfer, and syntactic parsing and discourse parsing are used as the source tasks for cross-task transfer. The results display the noteworthy effectiveness of the presented methods, with notable performance improvements over a baseline model without external knowledge.

An extensive analysis of individual features and transfer sources, and their specific impacts on the performance, underline the vital importance of choosing appropriate external knowledge sources. Moreover, the effectiveness of syntactic and discourse information varies across feature-based and transfer-based approaches, underscoring the different ways these types of knowledge can be integrated. These results not only offer valuable insights into the integration of external knowledge for argument mining, but also lay a solid foundation for future research exploring novel knowledge sources and integration strategies.

6.3 Chapter summary