Transfer or Translate? Argument Mining in Arabic with No Native Annotations

Sara Nabhani **Khalid Al-Khatib**
University of Groningen
{s.nabhani, khalid.alkhatib}@rug.nl

Abstract

Argument mining for Arabic remains underexplored, largely due to the scarcity of annotated corpora. To address this gap, we examine the effectiveness of cross-lingual transfer from English. Using the English Persuasive Essays (PE) corpus, annotated with argumentative components (Major Claim, Claim, and Premise), we explore several transfer strategies: training encoder-based multilingual and monolingual models on English data, machine-translated Arabic data, and their combination. We further assess the impact of annotation noise introduced during translation by manually correcting portions of the projected training data. In addition, we investigate the potential of prompting large language models (LLMs) for the task. Experiments on a manually corrected Arabic test set show that monolingual models trained on translated data achieve the strongest performance, with further improvements from small-scale manual correction of training examples.

1 Introduction

Argument mining is a subfield of natural language processing (NLP) concerned with the automatic identification of argumentative structures in text. These structures typically comprise components such as claims, premises, and major claims, which together form the backbone of rational discourse (Cabrio and Villata, 2018). Beyond its theoretical importance, argument mining has practical applications in domains such as education, online debate, misinformation detection, and policy analysis. Despite recent advances in neural methods and LLMs, research in argument mining has focused mainly on high-resource languages such as English (Li et al., 2025). In contrast, Arabic remains underexplored, largely due to the scarcity of annotated data. This gap limits the development of effective tools for argument analysis in Arabic-speaking contexts. Creating high-quality annotated resources for argument mining is both costly and time-consuming, especially in low-resource settings. A common strategy to address this challenge is to leverage existing English argumentation datasets through cross-lingual transfer or translation-based methods. Yet, the effectiveness of these approaches for a linguistically rich and structurally diverse language like Arabic remains an open question. In this paper, we investigate the feasibility of argument mining in Arabic by leveraging existing English resources. Our focus is on span-level argument component identification, which we formulate as a sequence labeling task using BIO-tagged annotations. Using the English Persuasive Essays corpus (Stab and Gurevych, 2017) as our source dataset, we evaluate four strategies:

- **Zero-Shot Multilingual:** Applying a multilingual model trained only on English directly to Arabic without adaptation.
- **Translate-Train Multilingual:** Training a multilingual model on a combination of English and translated Arabic data.
- **Translate-Train Monolingual:** Translating English training data into Arabic and training an Arabic model on the translated data.
- **LLM-Based Inference:** Prompting large language models to identify argument components in Arabic in a zero-shot setting.

We also conducted a small-scale annotation correction study to evaluate the impact of improving label quality in translated data. Our findings show that the Translate-Train Monolingual approach significantly outperforms alternative methods, and that even limited manual correction of projected labels yields substantial performance gains.

These findings highlight the effectiveness of translation-based modeling with minimal human supervision in addressing resource bottlenecks, while also emphasizing the need for high-quality,

Proceedings of The Third Arabic Natural Language Processing Conference, pages 407–416
November 8-9, 2025 ©2025 Association for Computational Linguistics

Arabic-specific argumentation corpora to support the development of more accurate and generalizable argument mining systems.

All the resources developed in this paper are available online.[1]

2 Related Work

Argument mining, the automatic analysis of argumentative structures in text, has advanced considerably in high-resource languages like English, supported by abundant annotated corpora and powerful models. However, research on low-resource languages such as Arabic remains limited due to scarce datasets and tools. To address this, recent efforts have started to build foundational resources for Arabic argument mining, while parallel work has explored cross-lingual transfer and the use of LLMs as potential solutions to the data bottleneck. In the following subsections, we review prior work in three key areas: Arabic argument mining, cross-lingual argument mining, and the application of LLMs to argument mining tasks.

Arabic Argument Mining A recent initiative in Arabic argument mining is Munazarat 1.0, a speech-based corpus comprising over 50 hours of transcribed MSA debates from QatarDebate tournaments, designed to support tasks such as debate strategy analysis and argumentation mining (Khader et al., 2024). Another notable effort is the 'Arabic Argumentative Debate' Corpus (Al-Sharafi et al., 2025), which applies a Toulmin-inspired, multi-dimensional scheme to label argumentative structures in debate transcripts. While both datasets are valuable, Munazarat 1.0 does not provide annotations for argumentative structure, whereas the Arabic Argumentative Debate Corpus focuses on higher-level rhetorical units.

Cross-Lingual Argument Mining Cross-lingual methods have been explored as a promising solution to the lack of annotated resources in low-resource languages. Eger et al. (2018) conducted one of the earliest studies in this space, introducing direct transfer and annotation projection techniques for argument mining between English and German. Their findings showed that translating English data and projecting annotations onto the target language could yield competitive results even without target-language supervision. Later work by Toledo-Ronen et al. (2020) confirmed the potential of such translation-based methods, showing that models like multilingual BERT can learn argumentative structures through machine-translated training examples, though performance declines somewhat when key language-specific nuances are lost in translation. Recent studies have shown that argument mining behaves differently from other sequence labeling tasks. Yeginbergen et al. (2024) tested several strategies in medical abstracts and found that translating data worked better than directly applying multilingual models. In the education domain, Ding et al. (2024) studied student essays written by English L1, English L2, and German learners. They found that differences in writing style and task type had a stronger effect on transfer performance than language alone.

In this paper, we follow a similar methodology to evaluate cross-lingual argument mining for Arabic. Specifically, we use English argumentation data, translate it into Arabic, and project the original annotations using word alignment tools. We compare this with other strategies including zero-shot transfer and training monolingual models on translated Arabic data. To our knowledge, this is the first study to systematically evaluate these approaches for Arabic argument mining.

Large Language Models for Argument Mining Recent studies have shown that LLMs can be highly effective for various argument mining tasks. A comprehensive survey by Li et al. (2025) outlines how LLMs, through prompt engineering, in-context learning, and chain-of-thought reasoning, can perform component identification and relation extraction. Gorur et al. (2024) demonstrated that open-source LLMs like Llama and Mistral can significantly outperform RoBERTa-based baselines on relation-based argument mining through careful prompting strategies. Meanwhile, Chen et al. (2024) evaluated models such as GPT, Flan, and Llama across several argument mining and generation datasets, finding strong performance even in zero- and few-shot settings. These promising results indicate that LLMs can handle both argument structure identification and relational reasoning. However, research in this area has focused almost exclusively on English. To our knowledge, little work has examined LLMs' zero-shot or few-shot performance on structured argument mining in low-resource languages such as Arabic. This is a gap our study seeks to address.

[1] https://github.com/saranabhani/ar-am-transfer

3 Data

Our main English resource is the Persuasive Essays (PE) corpus introduced by Stab and Gurevych (2017). This widely used dataset is annotated according to Freeman's theory of argumentation, which offers a simple yet generalizable framework. Prior work has demonstrated its utility for cross-lingual argument mining, showing that models trained on English can be adapted to low-resource languages (Eger et al., 2018). Building on this foundation, we investigate the extent to which English data can support argument mining in Arabic.

The PE corpus contains 402 English essays collected from `essayforum.com`. Each essay is paired with a description of the writing prompt to which it responds. The essays are segmented into paragraphs, and in our setup, each paragraph is treated as a separate data instance.

Each paragraph is annotated at the token level using the BIO (Begin, Inside, Outside) labeling scheme, where each token is tagged according to whether it is part of a **Major Claim**, **Claim**, or **Premise**:

- **Major Claim:** The central thesis or main argument of the essay.
- **Claim:** A proposition that supports and develops the Major Claim.
- **Premise:** A justification or evidence used to substantiate a Claim.

An example of an annotated essay segment from the PE corpus is shown in Figure 1. The dataset is already split into training and test sets, which we use as provided. Summary statistics for the corpus are shown in Table 1.

Statistic	Train	Test	Total
# Essays	322	80	402
# Paragraphs	1,786	449	2,235
# Tokens	118,645	29,537	148,182
Major Claim	598	153	751
Claim	1,202	304	1,506
Premise	3,023	809	3,832

Table 1: Statistics of the PE corpus across training and test splits.

4 Methodology

To address the cross-lingual challenge in Arabic argument mining, we experiment with five main approaches, grouped into three broad categories: Cross-Lingual Transfer, Translation-Based Training, and Large Language Models.

4.1 Cross-Lingual Transfer

We begin with a zero-shot setup where a multilingual model trained only on English data is applied directly to Arabic texts.

Multilingual Zero-Shot (EN) We train a multilingual model on the original English training data from the PE corpus. The model is then applied directly to Arabic texts without exposure to Arabic during training. This tests the model's ability to transfer argumentation knowledge across languages in a zero-shot setting.

4.2 Translation-Based Training

We investigate whether training on Arabic translations of the English corpus can enhance the performance. We test both multilingual and monolingual models under this setting.

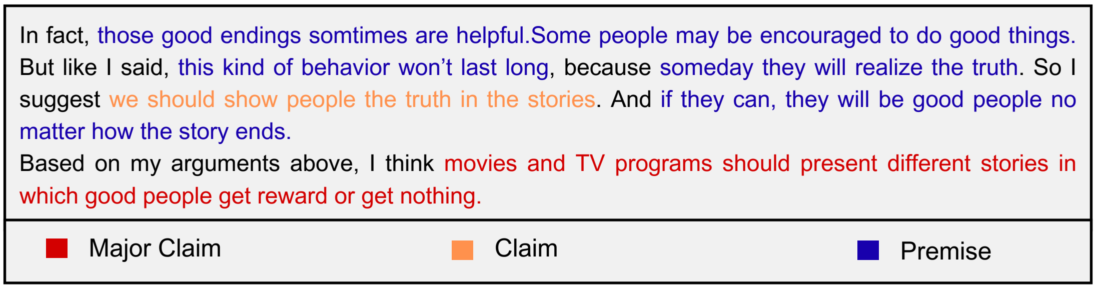
Multilingual Translate-Train (AR) We translate the English training data into Arabic and use it to train a multilingual model. This exposes the model to Arabic text during training, while still leveraging its multilingual capabilities.

Multilingual Combined Training (AR + EN) We train a multilingual model on a combination of the original English data and its Arabic translation. This setup allows the model to learn from both languages at once and potentially align representations across them more effectively.

Monolingual Translate-Train (AR) In this setting, we train a monolingual Arabic model using only the translated Arabic data. Unlike the previous two approaches, the model is not multilingual and is specialized in Arabic, which may help capture language-specific features more effectively.

4.3 Large Language Models Prompting

We also evaluate LLMs in a zero-shot setting. These models are prompted directly with Arabic task descriptions and examples, without any fine-tuning. This allows us to assess the out-of-the-box capabilities of general-purpose LLMs for Arabic argument mining.



In fact, those good endings somtimes are helpful.Some people may be encouraged to do good things. But like I said, this kind of behavior won't last long, because someday they will realize the truth. So I suggest we should show people the truth in the stories. And if they can, they will be good people no matter how the story ends.

Based on my arguments above, I think movies and TV programs should present different stories in which good people get reward or get nothing.

🟥 Major Claim 🟧 Claim 🟦 Premise

Figure 1: Example paragraph from the PE corpus

5 Experimental Setup

Building on the approaches outlined in the Methodology section, this section presents the experimental setup for evaluating Arabic argument mining using English resources. We describe the task formulation, model architecture, and training configurations, as well as the translation and annotation projection process and the evaluation setup, including a study on the impact of manual annotation correction.

5.1 Encoder-Based Models

This subsection outlines the experimental setup used to fine-tune encoder-based models for the task.

Model Architecture We formulate the task of argument mining as a sequence labeling problem, where the objective is to detect and classify contiguous spans of text corresponding to argument components. This formulation is supported by the structure of the PE corpus, which is annotated at the token level using the BIO tagging scheme.

Our architecture builds on prior work such as Eger et al. (2018), which employed BiLSTM-CRF models for argument component identification. In our setup, we replace the recurrent encoder with a transformer-based model to better capture long-range dependencies and contextual information. The overall model comprises three main components:

1. A pre-trained transformer encoder that processes tokenized input sequences
2. A token classification layer that produces label logits
3. A CRF layer that models label dependencies and ensures consistent label sequences

This architecture is used consistently across all fine-tuned experiments, with the primary difference being the choice of a pre-trained language model as the encoder, depending on the language and method used.

Models Used We experiment with both multilingual and monolingual transformer models, depending on the approach:

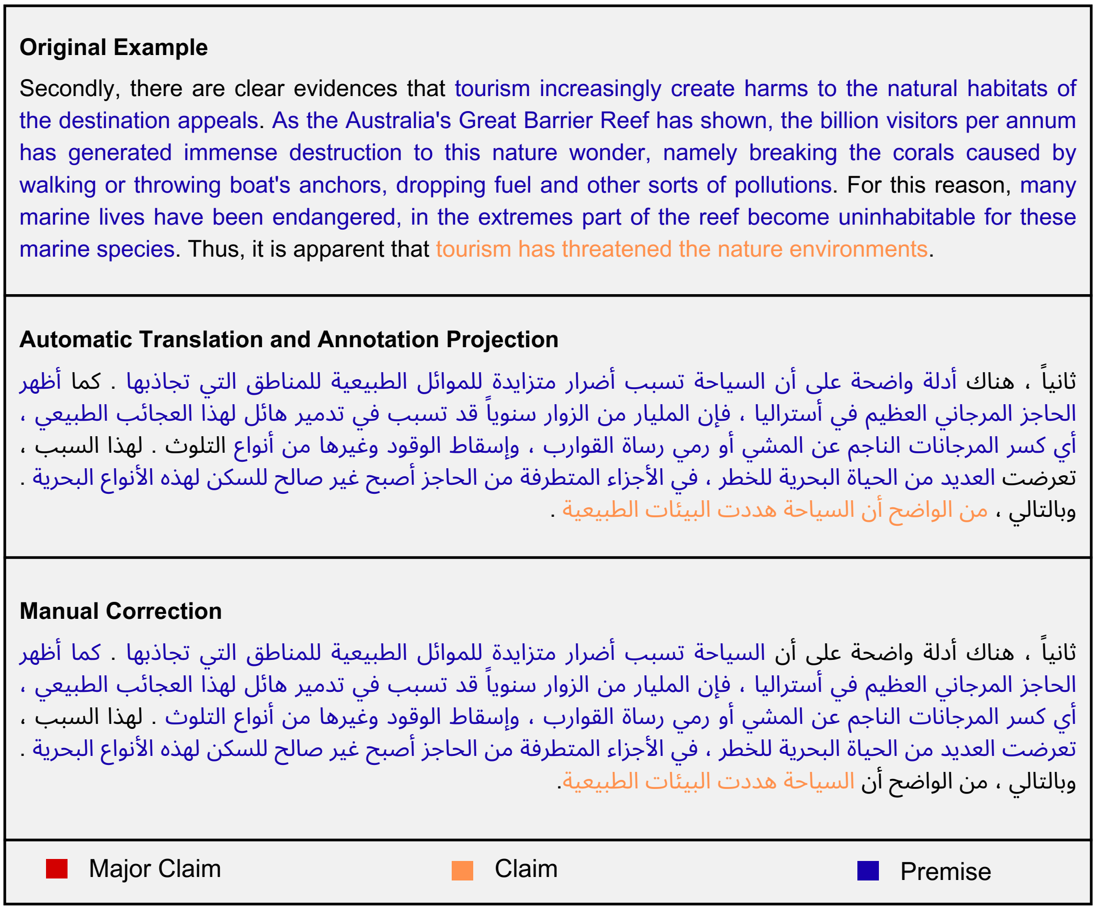
- **Multilingual Approaches:** We use XLM-RoBERTa-large (Conneau et al., 2019), a transformer model trained on 100 languages. Its strong cross-lingual capabilities make it suitable for both zero-shot and translation-based multilingual experiments.
- **Monolingual Approach:** For training directly on Arabic data, we use AraBERTv2 (Antoun et al.), a BERT-based model pre-trained specifically on large-scale Arabic corpora.

Training Configuration All models are fine-tuned using consistent hyperparameters, shown in Table 2.

Hyperparameter	Value
Max sequence length	256 tokens
Batch size	16
Epochs	100
Learning rate	3×10^{-5}
Weight decay	0.01
Warmup steps	100
Optimizer	AdamW

Table 2: Hyperparameters used for models fine-tuning.

Translation and Annotation Projection To create Arabic data for training and testing, we translate the English PE dataset using the NLLB model (No Language Left Behind) (Costa-jussà et al., 2022). We project the English token-level annotations onto the Arabic translation using FastAlign (Dyer et al., 2013), a widely used word alignment tool.



Original Example

Secondly, there are clear evidences that tourism increasingly create harms to the natural habitats of the destination appeals. As the Australia's Great Barrier Reef has shown, the billion visitors per annum has generated immense destruction to this nature wonder, namely breaking the corals caused by walking or throwing boat's anchors, dropping fuel and other sorts of pollutions. For this reason, many marine lives have been endangered, in the extremes part of the reef become uninhabitable for these marine species. Thus, it is apparent that tourism has threatened the nature environments.

Automatic Translation and Annotation Projection

ثانياً ، هناك أدلة واضحة على أن السياحة تسبب أضرار متزايدة للموائل الطبيعية للمناطق التي تجاذبها . كما أظهر
الحاجز المرجاني العظيم في أستراليا ، فإن المليار من الزوار سنوياً قد تسبب في تدمير هائل لهذا العجائب الطبيعي ،
أي كسر المرجانات الناجم عن المشي أو رمي رساة القوارب ، وإسقاط الوقود وغيرها من أنواع التلوث . لهذا السبب ،
تعرضت العديد من الحياة البحرية للخطر ، في الأجزاء المتطرفة من الحاجز أصبح غير صالح للسكن لهذه الأنواع البحرية .
وبالتالي ، من الواضح أن السياحة هددت البيئات الطبيعية .

Manual Correction

ثانياً ، هناك أدلة واضحة على أن السياحة تسبب أضرار متزايدة للموائل الطبيعية للمناطق التي تجاذبها . كما أظهر
الحاجز المرجاني العظيم في أستراليا ، فإن المليار من الزوار سنوياً قد تسبب في تدمير هائل لهذا العجائب الطبيعي ،
أي كسر المرجانات الناجم عن المشي أو رمي رساة القوارب ، وإسقاط الوقود وغيرها من أنواع التلوث . لهذا السبب ،
تعرضت العديد من الحياة البحرية للخطر ، في الأجزاء المتطرفة من الحاجز أصبح غير صالح للسكن لهذه الأنواع البحرية .
وبالتالي ، من الواضح أن السياحة هددت البيئات الطبيعية.

🟥 Major Claim 🟧 Claim 🟦 Premise

Figure 2: Example of one paragraph from the PE corpus in three forms: (a) the original English paragraph, (b) the automatically translated Arabic version with projected labels, and (c) the manually corrected Arabic version

5.2 LLM-Based Inference

In addition to the supervised systems, we test whether LLMs can recognize Arabic argument components without any task-specific fine-tuning. We use Llama 3.1 70B and Claude 3.5 Haiku in a zero-shot setting,[2] prompting the models and parsing their raw text output.

Prompt Design Each model receives a short *system* prompt that defines the task and a *user* prompt that provides the rules together with the paragraph to be labelled. Although the `{text}` placeholder is replaced with an Arabic paragraph at inference time, the prompts themselves are written in English because prior studies (Kmainasi et al., 2024) and our preliminary experiments indicate better performance when prompting in English rather than Arabic. The prompts are shown in Table 3.

5.3 Annotation and Translation Quality Study

Annotation Quality Annotation projection using alignment tools like FastAlign can introduce errors, especially for longer spans or less literal translations. To examine the effect of these errors, we manually reviewed a subset of the projected training annotations, corrected the identified errors, and measured the resulting change in model performance. The review was carried out in four phases of 100 paragraphs each, 400 in total. Paragraph- and token-level error rates, both relative to the reviewed subset and to the full training set, are summarized in Table 4.

These corrections were applied to translation-based experiments and allow us to evaluate how data quality influences learning outcomes. An example is shown in Figure 2.

Translation Quality We also reviewed 100 paragraphs for translation errors. The review revealed different types of errors, including morphological mistakes (e.g., “تَحَسَّنت ـ حَسَّنت” - “*was improved - improved*”), wrong word choices

[2] We also experimented with Fanar, an Arabic LLM (`https://huggingface.co/QCRI/Fanar-1-9B`), but observed very low output quality.

Prompt role	Content
System	You are a precise information extraction assistant. Your job is to identify and extract argumentative components from input text. These components include Major Claims, Claims, and Premises. Rules: - Extract spans exactly as they appear in the input — no rewriting. - A span can have only one label. - Spans must not overlap. - If there are no spans to extract, return nothing. - For each valid span, write the label on one line and the exact span on the next.
User	Your task is to extract any spans from the following text that represent a “Major Claim”, “Claim”, or “Premise”, if they exist. - Do not rephrase or alter the spans; extract them exactly as they appear. - Spans must not overlap. - Each span must have only one label. - If no spans exist in the text, do not output anything. Text: “{text}”

Table 3: System and user prompts supplied to Llama 3.1 70B and Claude 3.5 Haiku. The placeholder `{text}` is replaced with an Arabic paragraph at inference time.

Phase	out of reviewed	out of training
100	69%	3.9%
200	69%	7.7%
300	65%	10.9%
400	64%	14.4%

(a) **Paragraph-Level:** The proportion of paragraphs with at least one error among the manually reviewed paragraphs.

Phase	out of reviewed	out of training
100	16.2%	0.9%
200	15.4%	1.7%
300	13.6%	2.2%
400	13.2%	2.9%

(b) **Token-Level:** The proportion of tokens assigned wrong label among the tokens of the manually reviewed paragraphs.

Table 4: Error rates reported for each of the four review phases (100 paragraphs per phase; 400 total). For each phase, we show percentages relative to the reviewed subset and relative to the full training set.

(e.g., “تَشْريع ـ تُسَرِعْ” - “*acceleration - accelerates*”), literal translations of idioms (e.g., “*piece of cake*”), and minor spelling inconsistencies (e.g., “لدره ـ لدرجة” - “*to the extent*”). However, only 0.6% of the tokens in the reviewed sample contained translation errors. This rate is very low compared to the annotation error rates (see Table 4), and most of the identified errors were minor, with little to no impact on sentence meaning or structure. Due to this low error rate and its limited impact on data quality and argumentative label accuracy, we focus our correction study on annotation errors only.

5.4 Evaluation Setup

The evaluation is performed on the Arabic version of the PE test set. We translate the English test data using NLLB and project the original annotations using FastAlign. To ensure label consistency and fairness in evaluation, we manually review and correct the projected labels in the test set. We report precision, recall, and micro F1-score for each experiment. For translation-based approaches, we also investigate how annotation quality affects performance (Section 5.3).

6 Results

This section presents the results of the study, beginning with the evaluation outcomes and followed by an analysis of errors.

6.1 Evaluation Results

In this subsection, we report the results for each experimental setting described in Section 5.

The results are presented in two tables. Table 5 reports the performance of all models, both the encoder-based and the zero-shot large language models, using only the automatically projected data. Table 6 focuses on the annotation quality study showing how manual correction of a subset of the projected training data affects the performance of the fine-tuned models.

Model	P	R	F1
Multi-Ling. EN	0.009	0.001	0.003
Multi-Ling. AR	0.102	0.085	0.093
Multi-Ling. EN+AR	0.007	0.111	0.013
Mono-Ling. AR	**0.239**	**0.265**	**0.251**
Llama 3.1 70B	0.054	0.033	0.041
Claude 3.5 Haiku	0.149	0.085	0.108

Table 5: Performance of all models using the English PE training data and the Arabic translated data with **no manual correction**. Includes fine-tuned supervised models and zero-shot LLMs. **P** = Precision, **R** = Recall, **F1** = F1 score. All models are evaluated on the same manually corrected Arabic PE test set.

Model	#Rev	P	R	F1
Multi-Ling. AR	0	0.102	0.085	0.093
	100	0.100	0.090	0.095
	200	0.120	0.111	0.116
	300	0.093	0.096	0.094
	400	**0.137**	**0.119**	**0.128**
Multi-Ling. EN+AR	0	0.007	0.111	0.013
	100	0.136	0.107	0.120
	200	**0.154**	**0.129**	**0.140**
	300	0.142	0.116	0.128
	400	0.146	0.124	0.134
Mono-Ling. AR	0	0.239	0.265	0.251
	100	0.263	0.295	0.278
	200	0.309	0.340	0.324
	300	0.310	0.355	0.331
	400	**0.331**	**0.372**	**0.351**

Table 6: Effect of manual annotation correction on fine-tuned models. **P** = Precision, **R** = Recall, **F1** = F1 score. **#Rev** indicates the number of manually reviewed training examples (0, 100, 200, 300, or 400). All models are evaluated on the same manually corrected Arabic PE test set.

The results highlight the importance of both model choice and training data quality in cross-lingual Arabic argument mining. The monolingual model (AraBERT), trained on translated Arabic data, achieves the highest performance across all settings. Its F1 score improves steadily as more manually corrected training data is introduced, reaching 0.351 with 400 reviewed examples. These results confirm the findings by Yeginbergen et al. (2024) and extend them to Arabic.

Multilingual models, while generally lower-performing, also benefit from improved training labels. XLM-RoBERTa-large trained on both English and Arabic data performs poorly with uncorrected labels (F1 = 0.013), but improves significantly when 200 reviewed examples are included (F1 = 0.140). This shows that the quality of projected annotations has a strong effect on model performance, especially in cross-lingual setups.

The zero-shot multilingual model, trained only on English and evaluated directly on Arabic, performs very poorly (F1 = 0.003). This confirms that direct cross-lingual transfer is ineffective for this fine-grained sequence labeling task without any form of adaptation or supervision.

In the LLM setting, Claude 3.5 performs better than Llama 3.1, reaching an F1 score of 0.108. However, both models fall behind all encoder-based models, including those trained on noisy projected data. These results suggest that current large language models, while capable of some zero-shot generalization, still struggle with span labeling tasks in low-resource languages.

6.2 Error Analysis

The error analysis reveals that the model produces several types of errors: false positives, false negatives, misclassifications, and span boundary errors. Among these, boundary errors are the most frequent. In such cases, the model correctly identifies the argumentative type but predicts a span that is slightly longer or shorter than the gold annotation. This indicates that the model often locates the relevant segment in the text but struggles to precisely mark its start and end points. Misclassification errors are also common, where the predicted span matches the gold span but is assigned the wrong label. Less frequently, the model completely fails to detect a gold span (false negatives) or predicts a span that does not exist in the gold data (false positives).

For example, in the sentence:

“من الواضح أن السياحة هددت البيئات الطبيعية”

(“It is apparent that tourism has threatened the natural environments”)
the model predicted the entire sentence as a claim. However, in the gold annotation, only the part:

“السياحة هددت البيئات الطبيعية”

(“tourism has threatened the natural environments”)
is labeled as a claim. This illustrates a boundary error, where the model over-extends the span beyond the annotated target.

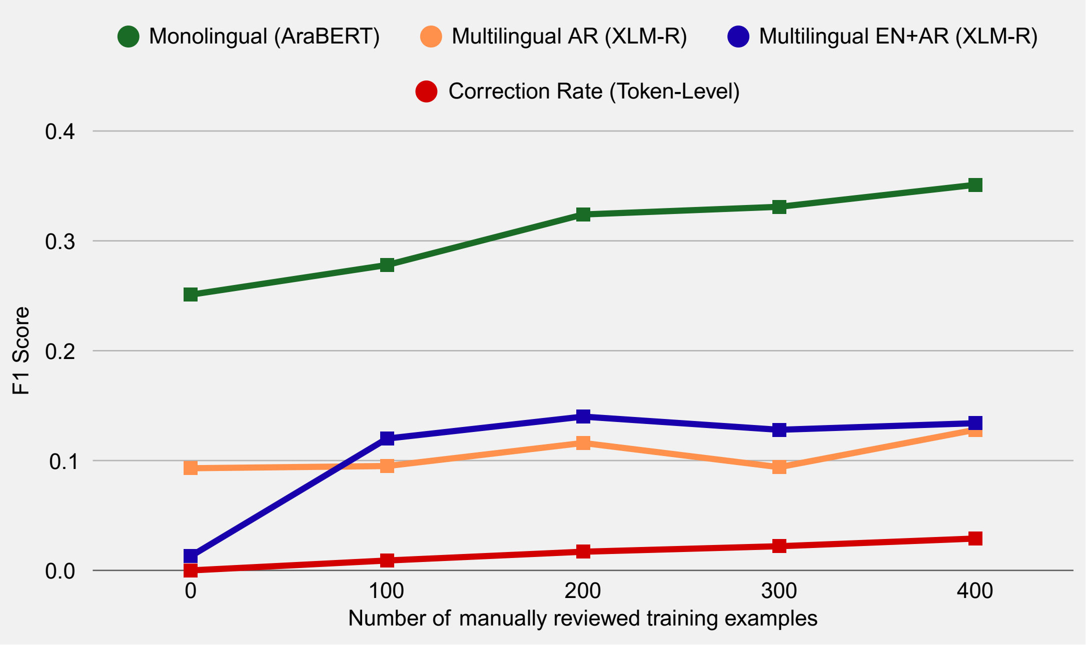


Figure 3: Model performance across different sizes of manually reviewed training instances

7 Discussion

In this section, we present the key findings derived from the results and outline the study's limitations.

7.1 Findings

Our findings highlight both the potential and the limitations of leveraging English resources to train Arabic argument mining models.

Multilingual pretraining is not sufficient by itself Despite its strong cross-lingual capabilities, XLM-RoBERTa performs poorly in the zero-shot setting. Even when trained on translated Arabic data, its performance remains lower than that of the monolingual AraBERT model. This suggests that multilingual models benefit from language coverage, but still require high-quality training data in the target language to succeed at structured tasks.

Translation-based training is effective, but sensitive to label quality Training on translated data works well when combined with a strong Arabic model. However, success depends on the quality of both the translation and the projected annotations. Tools like FastAlign are efficient for projection, but they struggle with more complex argument spans, particularly in longer or less literal translations.

LLMs are promising but not yet competitive LLMs like Claude 3.5 and Llama 3.1 showed some capacity for argument component detection in Arabic using zero-shot prompting. Claude 3.5, in particular, outperformed the zero-shot XLM-RoBERTa. However, both LLMs were outperformed by all encoder-based models trained on translated data, even without any manual correction. This suggests that while LLMs can serve as a baseline for sequence tagging in low-resource languages, they are not yet a reliable substitute for supervised training, particularly in token-level or span-based tasks such as argument mining.

Manual correction boosts performance, yet robust models require more manual annotation Introducing manual corrections to the projected training data had a clear and consistent effect, especially for AraBERT, as shown in Figure 3. Reviewing only 400 instances, that is less than 23% of the training set, led to meaningful gains in model performance. AraBERT's F1 increased by 10 percentage points, from 0.251 to 0.351. This demonstrates that even small-scale annotation can reduce noise and improve performance in projection-based pipelines. However, the improvements remain well below the level of a reliable model, showing that while limited correction is cost-effective, building a well-performing Arabic argument mining system ultimately requires a larger investment in high-quality annotation.

The need for an Arabic-specific corpus with human annotation The limitations of projection and translation point to the need for a high-quality, human-annotated Arabic argument mining dataset.

While translation-based training provides a strong starting point, and manual correction can boost performance, the results also show that these approaches have diminishing returns in the absence of clean, in-language supervision.

Creating a dedicated Arabic corpus with native annotations would allow models to learn the specific discourse, syntax, and argumentative structures used in Arabic. This resource would support more robust and accurate modeling and help close the gap between Arabic and high-resource languages in argument mining.

7.2 Limitations

Although our study demonstrates the potential of cross-lingual and translation-based methods for Arabic argument mining, some key limitations remain.

First, the reliance on automatic translation introduces the possibility of translation noise. Our preliminary analysis indicates that the NLLB model introduces very few errors with minimal impact on the translated data. However, we did not explicitly examine how even these limited errors might influence the performance of the downstream models. Future work could therefore investigate this connection more directly and also explore alternative translation models.

Second, our approach depends on annotation projection using FastAlign. While FastAlign provides a simple and efficient alignment strategy, it represents only one among several available approaches. More advanced techniques, such as neural alignment models or alignment methods that incorporate contextual embeddings, may yield more accurate projections of argumentative spans. Since our analysis does not compare different alignment strategies, we cannot fully assess how the choice of projection method impacts the quality of the annotation and the performance of the downstream model. Future work could explore alternative alignment techniques and systematically evaluate their effects on Arabic argument mining.

Third, our experiments with LLMs were limited. We only tested Llama 3.1 70B and Claude 3.5 Haiku in a zero-shot setting, without any task-specific training or examples. While this provides an initial sense of their ability to detect Arabic argument components, zero-shot prompting may not show their full potential. Using few-shot prompting, for example, may yield stronger and more reliable results. Future research could therefore extend our analysis by investigating a broader range of prompting and adaptation strategies to better understand the role of LLMs in Arabic argument mining.

8 Conclusion

This study investigated the feasibility of Arabic argument mining by leveraging English resources via cross-lingual and translation-based approaches.

We framed the task as span labeling, using the Persuasive Essays corpus to train and evaluate models across multiple strategies. Our experiments compared four approaches: Zero-Shot Multilingual, Translate-Train Monolingual, Translate-Train Multilingual, and LLM-Based Inference.

The results demonstrate that the Translate-Train Monolingual approach, which trains a dedicated Arabic model on translated English data, consistently outperforms all other methods. In contrast, multilingual models, even when exposed to Arabic, struggle to capture the linguistic and structural subtleties of argumentative discourse. Zero-shot and LLM-based inference settings showed limited performance, suggesting that neither multilingual generalization nor prompting alone is sufficient for this complex task.

Importantly, correcting even a small portion of projected training annotations yielded substantial performance gains in translation-based approaches, especially with the Monolingual Translate-Train approach. This finding emphasizes the value of high-quality annotation, even when applied at a limited scale.

Overall, our findings highlight both the potential and the limitations of leveraging English resources for Arabic argument mining. While translation and cross-lingual strategies offer a useful starting point, they cannot fully replace the need for carefully annotated Arabic resources. The results showed that modest manual correction is beneficial, yet not sufficient for a well-performing system. Building a robust Arabic argument mining system will therefore require sustained efforts to develop larger, higher-quality annotated corpora.

Acknowledgements

This work was partially supported by QD Fellowship award [QDRF-2025-02-020] from QatarDebate Center.

References

Abdul Gabbar Al-Sharafi, Mohammad Majed Khader, Mohamed Ahmed, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani, and Ali Al-Zawqari. 2025. A hybrid annotation model for arabic argumentative debate corpus. In *Arabic Language Processing*, Communications in Computer and Information Science, pages 97–113, Germany. Springer Science and Business Media Deutschland GmbH. Publisher Copyright: © The Author(s), under exclusive license to Springer Nature Switzerland AG 2025.; 8th International Conference on Arabic Language Processing, ICALP 2023 ; Conference date: 19-04-2024 Through 20-04-2024.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5427–5433. ACM.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, and Elahe Kalbassi et al. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yuning Ding, Julian Lohmann, Nils-Jonathan Schaller, Thorben Jansen, and Andrea Horbach. 2024. Transfer learning of argument mining in student essays. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 439–449, Mexico City, Mexico. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can large language models perform relation-based argument mining? *Preprint*, arXiv:2402.11243.

Mohammad M. Khader, AbdulGabbar Al-Sharafi, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani, and Ali Al-Zawqari. 2024. Munazarat 1.0: A corpus of arabic competitive debates. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 20–30, Torino, Italia. ELRA and ICCL.

Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2024. Native vs non-native language prompting: A comparative analysis. *Preprint*, arXiv:2409.07054.

Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025. Large language models in argument mining: A survey. *Preprint*, arXiv:2506.16383.

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

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. *Preprint*, arXiv:2010.06432.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024. Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11687–11699, Bangkok, Thailand. Association for Computational Linguistics.