A Cross-Domain Evaluation of Multimodal Argument Relation Identification

Oscar Morris

1 Introduction

2 Previous Work

3 Datasets

All datasets used in this project are available as corpora on AIFdb[1]. Using consistently annotated Argument Interchange Format (AIF) data allows many different datasets to be used and tested. The AIF Format [1] allows the annotation of argument data across all AM tasks, providing a platform for many different kinds of research.

Throughout the project two primary corpora have been considered: QT30 [2], a corpus consisting of 30 AIF annotated Question Time episodes, and a corpus of 9 AIF annotated Moral Maze episodes available on AIFdb.

3.1 Preprocessing

3.1.1 Argument Data

In order to use AIF data efficiently for ARI, it is useful to perform some preprocessing. This process produces a graph, where each node contains a locution, its related proposition, and the proposition's AIF identifier. This identifier corresponds to the audio data, allowing it to be easily loaded when required. Each edge in this graph corresponds to a relation between the propositions, one of RA (inference), MA (rephrase) or CA (conflict).

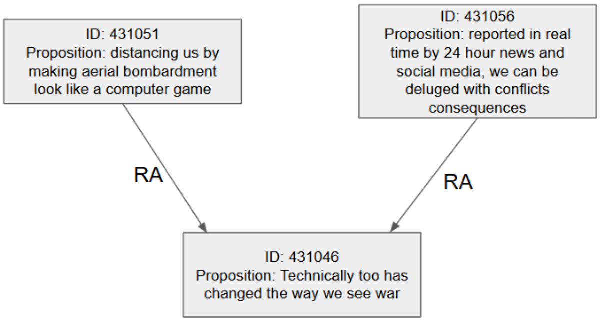
ID: 431051
Proposition: distancing us by making aerial bombardment look like a computer game
ID: 431056
Proposition: reported in real time by 24 hour news and social media, we can be deluged with conflicts consequences
RA
RA
ID: 431046
Proposition: Technically too has changed the way we see war

Figure 1: Example sub-graph.

Figure 1 shows an example sub-graph from the larger argument graph. Each node is truncated for brevity and only shows the node's ID, and the proposition. This sub-graph is taken from the Moral Maze episode on the 75th Anniversary of D-Day.

An example of the JSON structure used to store the argument data is shown in Listing 3.1.1.

```
{
    "id": 433407,
    "locution": "Matthew Taylor : she
    ↪  answered questions about
    ↪  norms and structures by
    ↪  talking about beliefs and
    ↪  campaigns and I think beliefs
    ↪  are different to norms and I
    ↪  think campaigns are different
    ↪  to social structures",
```

[1] https://corpora.aifdb.org/

```
    "proposition": "Nancy Sherman
     ↪  answered questions about
     ↪  norms and structures by
     ↪  talking about beliefs and
     ↪  campaigns and beliefs are
     ↪  different to norms and
     ↪  campaigns are different to
     ↪  social structures",
    "relations": [
        {
            "type": "CA",
            "to_node_id": 433393
        },
        {
            "type": "RA",
            "to_node_id": 433416
        }
    ],
}
```

An array of these JSON objects can then be used to create the node pairs required for the training and evaluation of the model.

3.1.2 Audio Data

The audio data first had to be downsampled from 44.1kHz to the 16kHz which is best accepted by the Wav2Vec2 transformer [3] among many others. This can easily be achieved using FFmpeg[2]. In the case of QT30, first audio had to be extracted from the video, and collapsed into a mono track before it could be downsampled, this was also easily achieved with FFmpeg.

Next, start and end times for each locution in the argument graph need to be found, to allow the audio to be split per-locution (and therefore per node). This can be achieved using PyTorch's forced alignment api[3].

Initially this was achieved by aligning each word in the transcript of the episode, producing start and end times for each word. A search can then be performed through this data to find the required locution. While this technique initially produced promising results, it was not robust enough to allow for errors in the transcripts or the crosstalk common in debates.

To solve this problem, the PyTorch forced alignment api is able to take wildcard tokens as input, therefore, each locution can be searched for individually. To achieve this, the partial transcript used as input to the forced aligner took the following form: `* {locution} *`.

Using this system allows the forced aligner to work well through crosstalk (since each locution's alignments are searched for independently of all others), and qualitatively seems to be more resilient to errors. Error resilience is helped since errors are less common in the locution texts as opposed to the transcripts. Using this system also allowed for confidence scores to be collected for analysis. In this section general analysis across all corpora is performed, with corpus specific analysis in the relevant section.

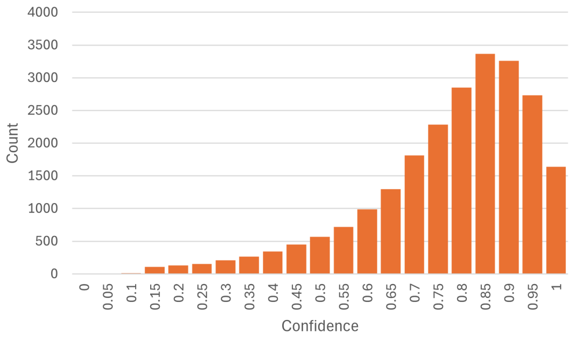

Figure 2: Confidence distribution across all corpora.

Figure 2 shows the distribution of confidence scores across both the QT30 corpus and all Moral Maze corpora. This distribution shows that the system can relatively confidently align the majority of locutions, with only approx. 8% of locutions with a confidence score less than 0.50.

In order to further analyse the performance

[2] https://ffmpeg.org/

[3] https://pytorch.org/audio/

of this system, locutions were selected at random and qualitatively analysed. Throughout this process, all locutions appeared correct, however, it was very challenging to accurately determine the accuracy of the system on locutions with confidence scores < 0.2. This shows that this method of aligning locutions with their corresponding audio is accurate.

3.1.3 Pair Creation

Finally, a set of node pairs and their relations can be generated in order to train a neural network. For related nodes this can be done trivially in that for each relation, the corresponding pair of nodes can be added to the set. When sampling unrelated nodes, however, things are more complex.

For this project, Short Context Sampling (SCS) is used as presented in [4]. Given the episodic structure of both QT30 and the Moral Maze corpora, a short context can be defined as the episode. This also allows the model to learn in a more realistic environment. Since the vast majority of node pairs have no relation, a number equal to that of Inference relations are generated.

3.2 QT30

The QT30 argument corpus [2] contains transcripts and argument annotations for 30 episodes of the BBC's Question Time, a series of televised topical debates across the United Kingdom. All episodes aired in 2020 and 2021. The corpus is split into 30 subcorpora, each spanning a single episode. This allows analysis of each episode individually, or combined as a single corpus.

Table 1 shows the distribution of each type of relation across QT30. Inference and Rephrase relations make up a total of 91.5% of the dataset, with Conflict relations being significantly less common, only making up 8.5% of the dataset. It is obvious that this is an unbalanced dataset, which will have to be considered during training.

Table 1: Disribution of propositional relations in QT30.

Relation Type	Count	Proportion (%)
Inference	5,761	51.4%
Conflict	947	8.5%
Rephrase	4,496	40.1%
Total	11,204	100%

Table 2: Mean confidence scores (μ) and standard deviation of confidence scores (σ) across each QT30 subcorpus.

Corpus Name	μ	σ
28May2020	0.76	0.16
4June2020	0.72	0.17
18June2020	0.76	0.16
30July2020	0.75	0.16
2September2020	0.78	0.15
22October2020	0.76	0.16
5November2020	0.77	0.17
19November2020	0.74	0.19
10December2020	0.77	0.16
14January2021	0.74	0.17
28January2021	0.70	0.17
18February2021	0.75	0.16
4March2021	0.76	0.16
18March2021	0.75	0.17
15April2021	0.75	0.15
29April2021	0.70	0.19
20May2021	0.76	0.17
27May2021	0.79	0.15
10June2021	0.75	0.17
24June2021	0.74	0.17
8July2021	0.72	0.17
22July2021	**0.44**	**0.28**
5August2021	0.76	0.17
19August2021	0.77	0.17
2September2021	0.78	0.16
16September2021	0.77	0.16
30September2021	**0.24**	**0.08**
14October2021	0.75	0.19
28October2021	0.75	0.17
11November2021	0.78	0.16
QT30	0.75	0.17

Table 3: Disribution of propositional relations in QT30-MM.

Relation Type	Count	Proportion (%)
Inference	5,740	51%
Conflict	937	8.4%
Rephrase	4,479	40.6%
Total	11,156	100%

Table 2 shows the mean and standard deviation of the confidence scores across each of the QT30 subcorpora. The lowest two mean scores are shown in bold. Manually analysing samples in these episodes indicates a high error rate in the alignment of locutions. Because of this high error rate, it was decided to exclude these episodes from the corpus used for training. The excluded episodes are: 22July2021 and 30September2021. The rest of the episodes from QT30 will form its multimodal subcorpus (QT30-MM).

Similarly to the complete QT30 corpus, Inference and Rephrase make up the vast majority of the QT30-MM dataset, with the proportion of Conflict relations decreasing to 8.4%.

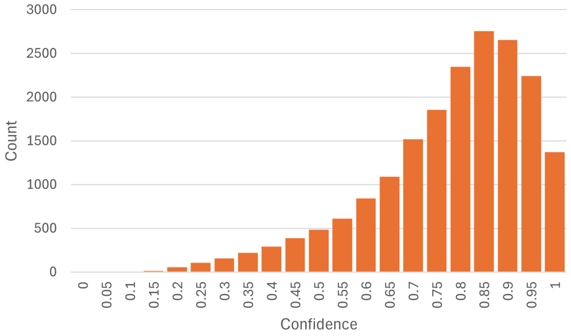

Figure 3: Confidence distribution across QT30-MM.

As can be seen in Figure 3 the distribution of confidence scores closely matches that shown in Figure 2. This still indicates that the audio alignments are calculated with high accuracy.

Table 4: Distribution of propositional relations after sampling non-related nodes.

Relation Type	Count	Proportion (%)
None	5,470	34%
Inference	5,470	34%
Conflict	937	6%
Rephrase	4,479	27%
Total	16,896	100%

3.3 Moral Maze

Similar to Question Time, the BBC's Moral Maze is a series of radio broadcast debates, with each episode focusing on a certain topic. Seven different Moral Maze episodes have been AIF annotated and made available on AIFdb. It is therefore these seven episodes, released from 2012 to 2019, which this project considers. Each episode focuses on a very different domain which allows for a robust, cross-domain analysis of any models trained on another corpus (e.g. QT30). The Moral Maze corpus contains data from nine different episodes: Banking (B), Empire (E), Money (M), Problem (P), Syria (S), Green Belt (G), D-Day (D) and Hypocrisy (H). Each episode consists of a debate focusing on a different topic, and hence has a different distribution of classes.

Table 5 shows the distribution of propositional relations across the Moral Maze corpus, after non-related pairs have been sampled. Comparing the corpus to QT30, a significantly lower proportion of the corpus is made up of Rephrase relations. It is possible that the differing formats of the debates has an impact here.

In Table 6 the mean and standard deviation of audio alignemnt confidence scores are compared across subcorpora. Generally the results match what is expected and are similar to those in QT30, the system does achieve unusually low scores when considering the D-Day subcorpus, the reason for this is unclear, however, manually analysing both random and low-confidence samples indicates they are gen-

Table 5: Distribution of propositional relations across the Moral Maze corpus.

Relation Type	None	Inference	Conflict	Rephrase	Total
B	132 (45%)	132 (45%)	24 (8%)	3 (1%)	291
E	151 (41%)	151 (41%)	39 (11%)	25 (7%)	366
M	255 (43%)	255 (43%)	29 (5%)	58 (10%)	597
P	236 (42%)	236 (42%)	40 (7%)	45 (8%)	557
S	181 (42%)	181 (42%)	63 (14%)	10 (2%)	435
G	301 (41%)	301 (41%)	46 (6%)	93 (13%)	741
D	72 (40%)	72 (40%)	7 (4%)	28 (16%)	179
H	207 (43%)	207 (43%)	23 (5%)	43 (9%)	480
Total	1,535 (42%)	1,535 (42%)	271 (7%)	305 (8%)	3,646

Table 6: Mean confidence scores (μ) and Standard Deviation of confidence scores (σ) across Moral Maze subcorpora.

Subcorpus	μ	σ
B	0.79	0.14
E	0.76	0.15
M	0.78	0.14
P	0.80	0.15
S	0.74	0.16
G	0.80	0.14
D	**0.50**	**0.34**
H	0.74	0.16

erally correct and so the subcorpus can be used for the cross-domain evaluation.

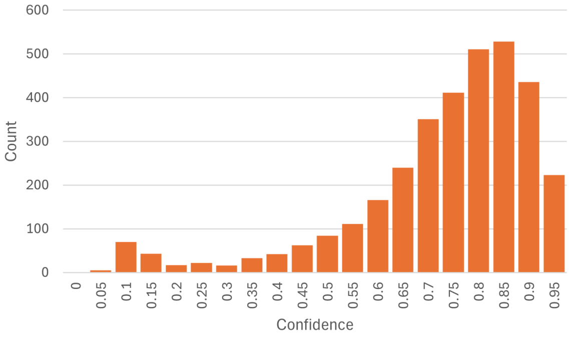

Figure 4: Confidence distribution across all Moral Maze subcorpora.

Figure 4 shows the distribution of audio alignment confidence scores across all Moral Maze episodes. This follows the expected pattern as shown in Figures 2 and 3. The secondary peak around a confidence of 0.10 is caused by the D-Day subcorpus.

4 Models

There are many different approaches when considering multimodal models. Generally they can be split into two categories: early fusion and late fusion. Early fusion models initially combine the features from each modality together, before being used as input to a single transformer model. Late fusion models use multiple transformers, one specialised in each modality. After being fed through each transformer, the hidden vectors are combined before being fed into the model's head. Initially, only late fusion models have been considered in this research.

This late fusion model uses RoBERTa-base [5] as its text transformer. To process the audio data, the model uses the Wav2Vec2-base audio transformer [3], having been pre-trained on 960 hours of dialogue[4]. The base models are used currently in order to increase the speed at which experiments can be conducted, however, they are used with a view to transitioning to their large variants towards the end of the project.

Figure 5 povides a visualisation of the primary model used at this point in the project. Similar

[4] https://huggingface.co/facebook/wav2vec2-base-960h

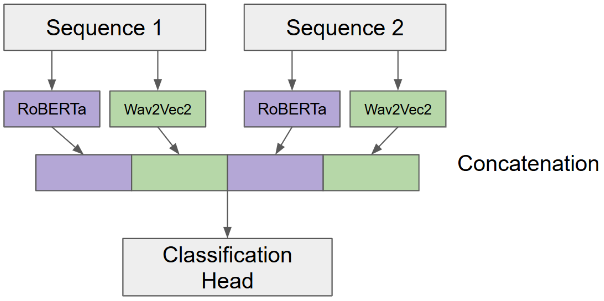

Figure 5: Concatenation model with late fusion.

models have been shown to achieve good multimodal performance while also being relatively simple to implement [6], [7]. It is for these reasons that this model has been implemented using the Huggingface[5] and PyTorch[6] python modules.

Some preliminary experiments have also been conducted using text-only models in an attempt to replicate results produced in [8]. For this the RoBERTa-base model is used and each sentence is concatenated before tokenisation, delimited by RoBERTa's special purpose token `</s>`.

5 Results

6 Conclusions

7 Future Work

References

[1] C. Chesñevar *et al.*, "Towards an argument interchange format," *The Knowledge Engineering Review*, vol. 21, no. 4, pp. 293–316, Dec. 2006, doi: 10.1017/S0269888906001044.

[2] A. Hautli-Janisz, Z. Kikteva, W. Siskou, K. Gorska, R. Becker, and C. Reed, "QT30: A Corpus of Argument and Conflict in Broadcast Debate," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Jun. 2022, pp. 3291–3300.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." arXiv, Oct. 2020. Accessed: Oct. 16, 2024. [Online]. Available: https://arxiv.org/abs/2006.11477

[4] R. Ruiz-Dolz, D. Gemechu, Z. Kikteva, and C. Reed, "Looking at the Unseen: Effective Sampling of Non-Related Propositions for Argument Mining," in *Proceedings of the 31st International Conference on Computational Linguistics*, Jan. 2025, pp. 2131–2143.

[5] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, Jul. 2019. doi: 10.48550/arXiv.1907.11692.

[6] E. Mancini, F. Ruggeri, A. Galassi, and P. Torroni, "Multimodal Argument Mining: A Case Study in Political Debates," in *Proceedings of the 9th Workshop on Argument Mining*, Oct. 2022, pp. 158–170.

[7] E. Mancini, F. Ruggeri, S. Colamonaco, A. Zecca, S. Marro, and P. Torroni, "MAMKit: A Comprehensive Multimodal Argument Mining Toolkit," in *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, Aug. 2024, pp. 69–82. doi: 10.18653/v1/2024.argmining-1.7.

[8] D. Gemechu, R. Ruiz-Dolz, and C. Reed, "ARIES: A General Benchmark for Argument Relation Identification," in *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, Aug. 2024, pp. 1–14. doi: 10.18653/v1/2024.argmining-1.1.

[5] https://huggingface.co/

[6] https://pytorch.org/