Predicting the Presence of Reasoning Markers in Argumentative Text

Jonathan Clayton Rob Gaizauskas
University of Sheffield
{jaclayton2,r.gaizauskas}@sheffield.ac.uk

Abstract

This paper proposes a novel task in Argument Mining, which we will refer to as *Reasoning Marker Prediction*. We reuse the popular Persuasive Essays Corpus (Stab and Gurevych, 2014). Instead of using this corpus for Argument Structure Parsing, we use a simple heuristic method to identify text spans which we can identify as reasoning markers. We propose baseline methods for predicting the presence of these reasoning markers automatically, and make a script to generate the data for the task publicly available [1].

1 Introduction

One key task within the field of argument mining (AM) is the generation of textual summaries of arguments (Fabbri et al., 2021; Bar-Haim et al., 2020). Significant work has been done on automatic extraction of argument components from argumentative text (see Lawrence and Reed, 2020 for a survey). However, research is still needed on how to use these extracted argument components to generate a fluent and readable textual summary.

One means to improve the coherence, and hence readability, of an argument summary is for the selected components which express the content of the argument to be connected using *reasoning markers*, rather than simply placing them adjacent to each other. Reasoning Markers are words and phrases such as “because", “therefore" or “in conclusion" which can be used to structure an argumentative piece of text, acting as the “glue" to hold a text together and make it more intelligible.

Figure 1 indicates how we might envision Reasoning Marker prediction being used in an argument summarisation pipeline. Such a pipeline could consist of argumentative components being extracted from a text, followed by selecting and ordering the most relevant components to form a summary. We concentrate on the final step of a proposed system like this; deciding where to insert reasoning markers to connect the selected argumentative components and produce a fluent text.

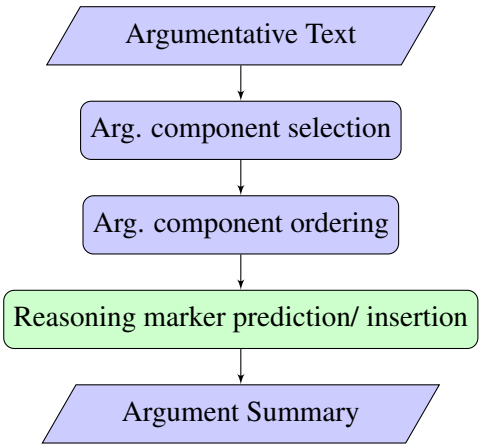

Figure 1: Our conceptualization for how Reasoning Marker Insertion could be used within an Argument Summarization pipeline.

1.1 Defining Reasoning Markers

Reasoning markers (RMs) are a proper subset of discourse markers (DMs), i.e. those words or phrases used in the organization of a spoken or written text.

Williams (2018) seems to be the first to have used the phrase “reasoning marker". However, the use of DMs in argumentative text has been noted since the notion of discourse marker was first introduced in Schiffrin (1987). The term RM excludes, for example, DMs which would typically be found in narrative text, such as “once upon a time", “eventually", or “suddenly".

RMs, specifically, are those discourse markers which are used to encode logical connections between claims and premises. The presence of RMs is argued to be positively correlated with the academic trustworthiness of a text (Williams, 2018).

We do not attempt to provide a rigorous definition of the notion of “reasoning marker" since we

[1] github.com/acidrobin/reasoning_marker_prediction

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 137–142
October 17, 2022.

believe this is a complex linguistic problem beyond the scope of this paper. Categories of discourse marker are notoriously difficult to define, and may be best conceptualized as "family resemblance" categories rather than categories definable by a list of formal features (Bordería, 2006).

Instead we sidestep the issue by assuming that whatever linguistic material can be used to connect together argument components counts as a Reasoning Marker. For our purposes we consider this definition satisfactory, since we are not aiming at formal linguistic correctness but generating a coherent and readable text.

1.2 Related Work

RMs have been used previously in argument mining as a feature for the identification of claims and premises, and the relations between them (Stab and Gurevych, 2014; Eckle-Kohler et al., 2015; Lawrence and Reed, 2015).

Malmi et al. (2017) build a large dataset for reasoning marker prediction, which they gather from English Wikipedia. Their dataset differs from ours in that it is not specifically aimed at argumentative text, and also uses sentence pairs instead of a short-essay context as in our work. Additionally, some authors have used discourse marker prediction as an auxiliary task for generating sentence embeddings (Sileo et al., 2019).

2 Corpus Creation

We use a simple heuristic method to identify RMs in an already existing corpus, taking advantage of existing annotations.

2.1 Persuasive Essay Corpus - Existing Annotation Scheme

The corpus which we choose to use for the extraction of Reasoning Markers is the Persuasive Essay Corpus (PEC) (Stab and Gurevych, 2017). PEC is a corpus of 402 persuasive essays on a variety of controversial topics. The corpus was annotated for the task of Argumentation Structure Parsing, i.e. identifying argumentative components within these essays and the links between them.

In order to extract Reasoning Markers from PEC, we use a heuristic rule-based method. We note that PEC comes pre-segmented into Argument Components (ACs). A BIO tagging schema is used to label each token as either belonging to an AC or not; and, if a token belongs to an AC, it is labelled as either a Claim, a MajorClaim (the claim that is the main topic of each essay) or a Premise.

[Furthermore , RM] [investing in art could bring employment opportunities and could end in return of capital occasionally CLAIM] . [The investment could be paid back through the values of the created works of art which as a matter of fact should be considered as national possessions PREMISE] [. To sum up , RM] [not only could investing in art be considered as wasting money at any kind PREMISE] [, but also RM] [it would enriches the culture of the society PREMISE] .

Figure 2: An essay fragment from PEC with our automatically generated RM annotations applied to it. Note we show annotated spans rather than tokens for readability. Tokens in spans labelled "RM" are originally labelled "O" in PEC.

Looking at an example from PEC in Figure 2, we can observe that some ACs are separated by RMs, while others are separated only by punctuation. This suggests that it may be possible to leverage this dataset for RM prediction.

2.2 Inferring Reasoning Markers

In order to identify RMs, we use a simple two-stage pipeline: (1) carry out sentence tokenization; (2) identify those segments within a sentence containing an AC (Claim, MajorClaim or Premise) but labelled with O tags, excluding segments consisting solely of a single punctuation character.

We observe that the vast majority of these O-labelled sentence fragments can be considered as either constituting or containing an RM. This should not be surprising for two reasons: (1) as just outlined, all of these sentence fragments come attached to ACs; (2) the essays originate from `essayforum.com`, a website consisting mostly of essays composed by high-school students or learners of English as a second language – educational contexts where students are rewarded for including RMs within texts.

2.3 Corpus Contents

We find our processed version of PEC contains a total of 7426 "potential RM" datapoints, where a potential RM datapoint occurs between each pair of adjacent ACs. The data is evenly balanced between the classes RM/No RM, as can be seen in Table 1.

The corpus contains a total of 1264 reasoning marker types. While this number seems large, it is somewhat artificially inflated by a number of minor

	RM	No RM	Total
Train	2726	2550	5276
Validation	346	287	633
Test	802	715	1517
Total	3874	3552	7426

Table 1: Numbers of samples found in train, validation and test sets.

variations on what are semantically very similar RM phrases, such as “To conclude, I definitely feel that", “To conclude, I strongly believe that", “To conclude, I want to say that", and a number of similar examples. Shorter RMs are also much more common than longer RMs, as shown in Table 2 and Figure 3. 39.8% of all RMs are only a single token long. As well as concurring with Zipf’s brevity law (Zipf, 1949), this reflects the length of RMs typically studied in the literature.

Reasoning Marker	Frequency in Corpus
“because"	195
“for example"	178
“therefore"	137
“however"	110
“moreover"	104

Table 2: The five most common reasoning markers appearing in PEC

The classification of some of the longer segments as RMs is somewhat dubious. For example, the following 25-token phrase would not be typically classified linguistically as a RM, but it seems to fulfil a similar function in context:

> “In conclusion, after analyzing the pros and cons of advertising, both of the views have strong support, but it is felt that... «conclusion»"

However, we refrain from filtering out discourse markers using linguistic criteria, since we treat this as an engineering task and mainly aim to add in appropriate connective material between ACs, whether or not they count as RMs in the strict sense.

2.4 The Corpus Processing Script

We release a script via our repository (`github.com/acidrobin/reasoning_marker_prediction`) which takes in the data provided in the PEC repository (`github.com/UKPLab/acl2017-neural_end2end_am`) and converts it into valid input for a language model. We describe the format of this input in Section 3.1.

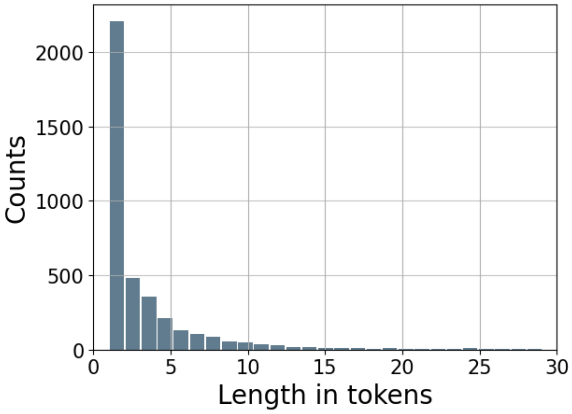

Figure 3: Counts of RM tokens in PEC by token length

3 Task and Baselines

Here we describe the task and the implementation of several pretrained language model baselines.

3.1 Reasoning Marker Prediction

In this task, we take a version of PEC with full essays represented as strings, with each essay replicated as many times as there are RMs in it. Each essay copy has a single gap where one RM may or may not appear. We then predict whether or not an RM should appear in this gap. All other RMs are included in this copy, but excluded in turn in other copies. This is a binary classification task, with two possible labels “True" (if an RM is present) and "False" (if an RM is not present).

Input: “furthermore investing in art could bring employment opportunities and could end in return of capital occasionally [RM] the investment could be paid...“
Output: “False"

Table 3: Input/output schema for RM prediction task.

We also add an additional test condition, which we denote +AC, in which we add special tokens to the input representing AC types: [claim], [major-claim], and [premise]. For example, a paragraph containing a premise and claim would have an input similar to “[premise] *premise text* [RM] [claim] *claim text*" – so that the [premise] and [claim] special tokens indicate the beginning of these components. We reason that in at least some cases this should help provide useful information to the model; e.g. RMs such as “in conclusion" are very common in the dataset before major claims.

3.2 Implementation of Baselines

Since the ratio of RMs to no-RMs is roughly 50/50, we use a random baseline where the probability of choosing “True" is set at 0.5.

We also use two large pretrained language models, BERT (Devlin et al., 2018) and T5 (Raffel et al., 2020). The two share fundamental similarities in that they are transformer models (Vaswani et al., 2017), however they differ in the specific pretraining regime that they follow (see Raffel et al. (2020) for details).

Implementations of the two models are taken from `huggingface.com` (Wolf et al., 2019). For both models, we used lower-cased text in the input, to prevent trivial classification using the case of the word following the potential RM location.

Due to its pretraining scheme, T5 benefits from a task-specific linguistic prompt being prepended to the input. We experiment with options: no prompt, “True or False:", and “Is there a reasoning marker? True or False:". We found the prompt “True or False:" gave the best result.

For the tokenizers of both models, we add in a special [RM] token which indicates a potential reasoning marker position. For the +AC test condition, we add [claim], [majorclaim], and [premise] special tokens to the BERT tokenizer. For the T5 model, we additionally add “true" and “false" as single tokens. To use the T5 model, which can generate free text, as a classifier, we generate only a single token at inference time and ignore all logits except those corresponding to the “true" and “false" tokens.

Appendix A contains further details of our training scheme.

4 Results

We evaluate our results using precision, recall and F1-score macro-averaged between the two classes. The random baseline, as might be anticipated, achieved an F1-score of 0.50.

Model Name	Precision	Recall	F1-Score
RandomBaseline	0.50	0.50	0.50
bert-base	**0.75**	**0.70**	**0.69**
bert-base+AC	0.73	0.66	0.64
t5-small	0.63	0.60	0.59

Table 4: Performance of the baseline and three models evaluated on the test set.

As Table 4 shows us, the best-performing model was the “vanilla" bert-base-uncased. Adding in the extra tokens to indicate the beginning of argument components lowered performance. Additionally, the T5 model underperformed compared to BERT. The reasons for this are unclear – one possible explanation that could be hypothesized is that the input to this task is closest to what was seen in pretraining by the bert-base model since it was all uncased. The T5 model used was cased since an uncased variant was not available on the web. The largest source of error for both models was over-prediction of reasoning markers.

5 Conclusions and Future Work

We have presented a new task which we believe is a useful subtask for generating summaries of argumentative text: reasoning marker prediction. We have released a script that can be used to generate our derived corpus from PEC, which supports this task. Additionally, we have shown it is possible to predict the presence or absence of an RM between two argumentative components at an above-chance level. Our baseline scores show this is a challenging task, with much room for improvement.

Of course we want not only to predict *that* an RM should occur but *what* the RM should be. In the future, we aim to work on using end-to-end models to generate an appropriate RM for a given context, instead of simply predicting whether or not an RM should appear.

Another aspect of this task which we have not explored is the sub-categorization of RMs. Multiple taxonomies of DMs have been developed that could be used for this task. See Knott’s (1996) taxonomy, and the development in Oates (2000).

However, it is likely that this would be a non-trivial task and require some expert labelling, due to the fact that there is not a one-to-one correspondence between DMs and their functions. A simple DM like “so" for example, has many different functions and can be used to provide justifications, for sequencing, or for expressing a purpose.

Nonetheless, since, as noted above, there are many RMs in this dataset that are more-or-less interchangeable, it may be sufficient to predict the category that a potential RM should belong to rather than attempting to generate one directly.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1], and by Amazon.

References

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*.

Salvador Pons Bordería. 2006. A functional approach to the study of discourse markers. In *Approaches to discourse particles*, pages 77–99. Brill.

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*.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

Alexander R Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *arXiv preprint arXiv:2106.00829*.

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

Alistair Knott. 1996. A data-driven methodology for motivating a set of coherence relations.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

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

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2017. Automatic prediction of discourse connectives. *arXiv preprint arXiv:1702.00992*.

Sarah Louise Oates. 2000. Multiple discourse marker occurrence: Creating hierarchies for natural language generation. In *Proceedings of the ANLP-NAACL 2000 Student Research Workshop*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Damien Sileo, Tim Van-De-Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. *arXiv preprint arXiv:1903.11850*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashley Williams. 2018. Using reasoning markers to select the more rigorous software practitioners' online content when searching for grey literature. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, pages 46–56.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort: an introduction to human ecology.

A Details of Training Scheme

All of our BERT and T5 models are pretrained and then fine-tuned on the task for a number of epochs chosen by early stopping, in the range of $[0 \,..\, 8]$. We used the uncased version of BERT base. We use the Adam optimizer (Kingma and Ba, 2014). The best learning rate is chosen by a grid search; for both models we explore the set $\{1e^{-4}, 5e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}\}$. For the BERT model, we found the optimal learning rate was $1e^{-5}$ and the best performance was achieved after 3 epochs of fine-tuning. For T5, a learning rate of $5e^{-6}$ and 5 epochs of fine-tuning were optimal.