Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models

Lawrence, John; Reed, Chris

Published in:
Proceedings of the 4th Workshop on Argument Mining

DOI:
10.18653/v1/W17-5105

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models

John Lawrence and **Chris Reed**
Centre for Argument Technology,
University of Dundee, UK

Abstract

This paper presents a method of extracting argumentative structure from natural language text. The approach presented is based on the way in which we understand an argument being made, not just from the words said, but from existing contextual knowledge and understanding of the broader issues. We leverage high-precision, low-recall techniques in order to automatically build a large corpus of inferential statements related to the text's topic. These statements are then used to produce a matrix representing the inferential relationship between different aspects of the topic. From this matrix, we are able to determine connectedness and directionality of inference between statements in the original text. By following this approach, we obtain results that compare favourably to those of other similar techniques to classify premise-conclusion pairs (with results 22 points above baseline), but without the requirement of large volumes of annotated, domain specific data.

1 Introduction

The continuing growth in the volume of data which we produce has driven efforts to unlock the wealth of information this data contains. Automatic techniques such as Opinion Mining and Sentiment Analysis (Liu, 2010) allow us to determine the views expressed in a piece of textual data, for example, whether a product review is positive or negative. Existing techniques struggle, however, to identify more complex structural relationships between concepts.

Argument Mining is the automatic identification of the argumentative structure contained within a piece of natural language text. By automatically identifying this structure and its associated premises and conclusions, we are able to tell not just *what* views are being expressed, but also *why* those particular views are held. Argument mining has recently been enjoying rapid growth, propelled by three drivers: first, the academic and commercial success of opinion mining and sentiment analysis techniques upon which argument mining builds; second, a strong commercial appetite for such technologies from companies such as IBM; and third, the development of infrastructure and tools for (Bex et al., 2013), and theoretical understanding of (Budzynska et al., 2014), argument structure in both monologue and dialogue.

The intuition underlying the work presented here is that there are rich and predictable thematic and lexical commonalities present in the expression of human reasoning, and that these commonalities can be identified in helping to extract the structure of reasoning. For example, in debates concerning abortion, arguments are carefully marshalled on both sides, with religious themes more typically appearing on one side, and feminist philosophy themes more typically on the other. For a debate on the construction of a new road, we may find environmental issues on one side and economic concerns on the other. If such generalisations are possible at a coarse scale, perhaps they are similarly possible at a more fine-grained scale.

These themes are represented in terms of both the topics discussed and the language used to express them: an anti-abortion stance is likely to cover, not just feminist philosophy themes in general, but to use specific terminology more frequently, perhaps mentioning 'choice' or 'freedom' more than views expressed on the other side. When humans hear such a debate, they understand the structure of the arguments being made not only based on the content of the argument itself, but

Proceedings of the 4th Workshop on Argument Mining, pages 39–48

on a broad general knowledge of the topic and the way in which such arguments are commonly presented.

The argument mining technique which we present in this paper takes the commonly occurring terms in the original text and then uses these terms to gather data from the web on the same topic. This large volume of additional data can be considered as contextual knowledge, and is processed to find pairs of text spans which have an inferential relationship. We then use these pairs to create premise-conclusion topic models, reflecting the ways in which one topic or phraseology is used to support another.

Previous work (Lawrence and Reed, 2015) has shown that discourse indicators such as *because* and *therefore* are very reliable predictors of argument structure. Unfortunately they are also rather rare, occurring with fewer than 10% of argumentative inference steps. With a high-precision/low-recall technique such as is provided by these indicators, it becomes possible to process large amounts of text to extract a dataset in which we can have high confidence. This dataset can be used to capture topical regularities in the argument structure which can then be exploited in analysing text which does not benefit from the presence of indicators.

2 Related Work

The majority of the work carried out to date in the field of argument mining, has used either a supervised learning approach (e.g. (Palau and Moens, 2009; Feng and Hirst, 2011; Stab and Gurevych, 2014)), or a linguistic rule-based approach ((Villalba and Saint-Dizier, 2012; Pallotta and Delmonte, 2011; Wyner et al., 2012)), to determine argumentative function. In both cases these efforts are limited by a lack of consistently annotated argument data. Whilst resources such as the Internet Argument Corpus (IAC) (Walker et al., 2012) and AIFdb (Lawrence et al., 2012), offer rapidly growing volumes of high quality argument analyses, they do not provide the large volumes of data required to train a robust classifier, particularly when considered in the context of a specific topic or domain.

Attempts have been made to mitigate this constraint by the automatic creation of argument corpora, however, the datasets produced are limited to very specific types of data. For example, in (Houngbo and Mercer, 2014), a straightforward feature of co-referring text using the word "this" is used to build a self-annotating corpus extracted from a large biomedical research paper dataset. This is achieved by collecting pairs of sequential sentences where the second sentence begins with "This method...", "This result...", or "This conclusion...", and then categorising the first sentence in each pair respectively as Method, Result or Conclusion sentences.

Similarly, in (Habernal and Gurevych, 2015), unsupervised features are developed for argument component identification which exploit clustering of unlabelled argumentative data from online debate portals. Al-Khatib et al. (2016) likewise leverage online debate portals, applying distant supervision to automatically create a large annotated corpus with argumentative and non-argumentative text segments from several domains.

Our approach to expanding the data available on the topic under discussion relies on the high precision identification of inferential relationships shown by the presence of discourse indicators. Discourse indicators are explicitly stated linguistic expressions of the relationship between statements (Webber et al., 2011), and, when present, can provide a clear indication of argumentative structure. For example, if we take the sentence "Britain should disarm because it would set a good example for other countries", then this can be split into two separate propositions "Britain should disarm" and "it [disarming] would set a good example for other countries". The presence of the word "because" between these two propositions clearly tells us that the second is a reason for the first.

Discourse indicators have been previously used as a component of argument mining techniques, for example in (Stab and Gurevych, 2014) indicators are used as a feature in multiclass classification of argument components, with each clause classified as a major claim, claim, premise or non-argumentative. Similar indicators are used in (Wyner et al., 2012), along with domain terminology (e.g. camera names and properties) to highlight potential argumentative sections of online product reviews. In (Eckle-Kohler et al., 2015) a German language corpus is annotated with arguments according to the common claim-premise model of argumentation and the connection between these annotated connections and the presence of discourse indicators (or discourse markers

as they are referred to here) is investigated. The results presented show that discourse markers are again important features for the discrimination of claims and premises in German as well as English language texts.

There are many different ways in which indicators can appear, and a wide range of relations which they can suggest (Knott, 1996). For automatic corpus construction, the ability to identify all of these connections is not relevant and we are able to concentrate solely on those indicators offering a very high chance of describing an inferential relationship.

Using discourse indicators to build such a corpus is supported by the work done in identifying implicit discourse relations, for example (Lin et al., 2009; Park and Cardie, 2012), where a range of relations labelled in the Penn Discourse TreeBank (Prasad et al., 2008), but not explicitly indicated, were identified using features from those relations where an explicit indicator did occur. These implicit relations were identified with accuracies of between 70-80% in one-vs-others tests, clearly suggesting that studying cases where indicators are present can give a strong indication of a relationship in those cases where they are omitted.

The relationship between the topics being expressed in a piece of text and the argumentative structure which it contains have been previously explored in (Lawrence et al., 2014), where a Latent Dirichlet Allocation (LDA) topic model is used to determine the topical similarity of consecutive propositions in a piece of text. The intuition is that if a proposition is similar to its predecessor then there exists some argumentative link between them, whereas if there is low similarity between a proposition and its predecessor, the author is going back to address a previously made point and, in this case, the proposition is compared to all those preceding it to determine whether they should be connected. Using this method a precision of 0.72, and recall of 0.77 are recorded when comparing the resulting structure to a manual analysis, however it should be noted that what is being identified here is merely that an inference relationship exists between two propositions, and no indication is given as to the direction of this inference.

3 Experimental Data

The data used in this paper is taken from a transcript of the BBC Radio 4 program *Moral Maze*[1]. Specifically, we look at the episode from July 4th 2012[2] on the morality of the banking system. Manual argumentative analysis was performed on the transcript, using the OVA+ (Online Visualisation of Argument) analysis tool (Janier et al., 2014) to create a series of argument maps capturing the structure using the Argument Interchange Format (AIF) (Chesñevar et al., 2006). A corpus containing the full manual analysis of the transcript can be found online at `http://corpora.aifdb.org/bankingsystem`. The corpus comprises 5,768 words, split across 327 propositions, with 128 inferential connections (premise/conclusion relations) between them.

Identifying the argumentative structure contained within a piece of text can be viewed as a two-step process: Firstly, identifying the individual units of discourse which the text contains (commonly referred to as 'Argumentative Discourse Units' or ADUs (Peldszus and Stede, 2013)); and then, determining the ways in which these propositions are connected.

Figure 1 shows the AIF compliant representation of a fragment of the Moral Maze dialogue. In this figure, the blue boxes represent individual ADUs, while the arrows show connections, and the diamonds detail the nature of these connections. In this case, the conclusion "I know bankers who behave absolutely splendidly" is supported by the individual premises "who are major benefactors", "who spend their Christmases manning soup kitchens", and "Think about Bill Gates and all the wonderful things that his money is doing".

We can see from this example that the broad concept of charitable works is being used to support the idea that bankers are good people. The knowledge that these premises are both thematically related and support the character of a group of people, whilst clear to a human analyst, is not explicitly indicated in the original text.

For our purposes, we are aiming to identify inferential connections between pairs of ADUs. Whilst a complete argument mining pipeline would require the automation of this segmentation, this is outside the scope of this paper, and the focus of much additional research within the argument mining field (Lawrence et al., 2014; Mad-

[1] `http://www.bbc.co.uk/programmes/b006qk11`

[2] `http://www.bbc.co.uk/programmes/b01kbj37`

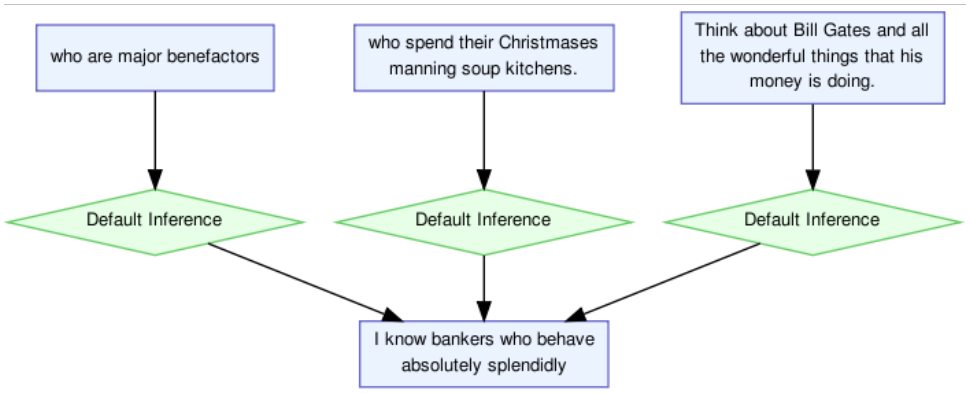

Figure 1: Fragment of Manually Analysed Argument Structure from the BBC Radio 4 program *Moral Maze*

nani et al., 2012; Saint-Dizier, 2012). As such, we use the same segmentation carried out for the manual analysis, and split the possible ADU pairs into those which are connected by an inferential relationship, and those which are not.

4 Implementation

An overview of the methodology used can be seen in Figure 2. Starting with raw, natural language text, manual segmentation is performed to split the text into ADUs. From here these segments are examined in order to find those unigrams and bigrams which occur most frequently throughout the text, giving an indication of the overall theme of the text which we are working with.

The next step is then to build a corpus of related documents by searching the web for those unigram and bigram terms identified as being indicative of the theme. From this extended corpus, we then extract sentences which contain an inferential relationship by searching for those discourse indicators which we have found to have the highest precision. This search results in a large collection of pairs of text fragments where one of the pair is a premise supporting the other.

Using these fragments as documents, we then generate a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model, and from this create a matrix showing the probability of support between each of the identified topics. By matching pairs of ADUs from the original text against the probabilities in this matrix, we are then able to determine the probability that there is an inferential relationship between them, and by thresholding these values, we can then categorise ADU pairs as being 'inferential' or 'non-inferential'.

An alternative approach would be to use the premise/conclusion dataset as training data for a supervised machine learning approach. This is limited by the fact that we only obtain positive examples, and, whilst techniques such as PU-learning (Learning from Positive and Unlabelled examples) (Liu et al., 2003) provide a way of dealing with only positively labelled data, we do not have sufficient quantities of unlabelled examples for these techniques to be applied. In future work, the ability to identify arbitrary ADUs in text could be used to extract large volumes of unlabelled examples, and such approaches may then become more suitable.

4.1 Obtaining Premise/Conclusion Pairs

The first step in the pipeline described above is to determine the overall theme of the text being analysed. This was performed by looking for those unigrams and bigrams which occur most frequently throughout the text. With the text previously segmented into ADUs, we calculated the number of unique ADUs in which each unigram or bigram appeared. This list is then sorted and filtered to remove common stop words. The resulting lists of terms can be seen in Table 1 and Table 2.

Having identified keywords describing the topic, a corpus of related documents was created by searching the web for combinations of these terms. The top ten terms of each kind were combined into search queries by taking all possible combinations of two and three unigrams as well as each bigram both on its own and paired with each unigram. Using these queries, the first 200 Google search results for each were compiled. After filtering the list of related documents to remove

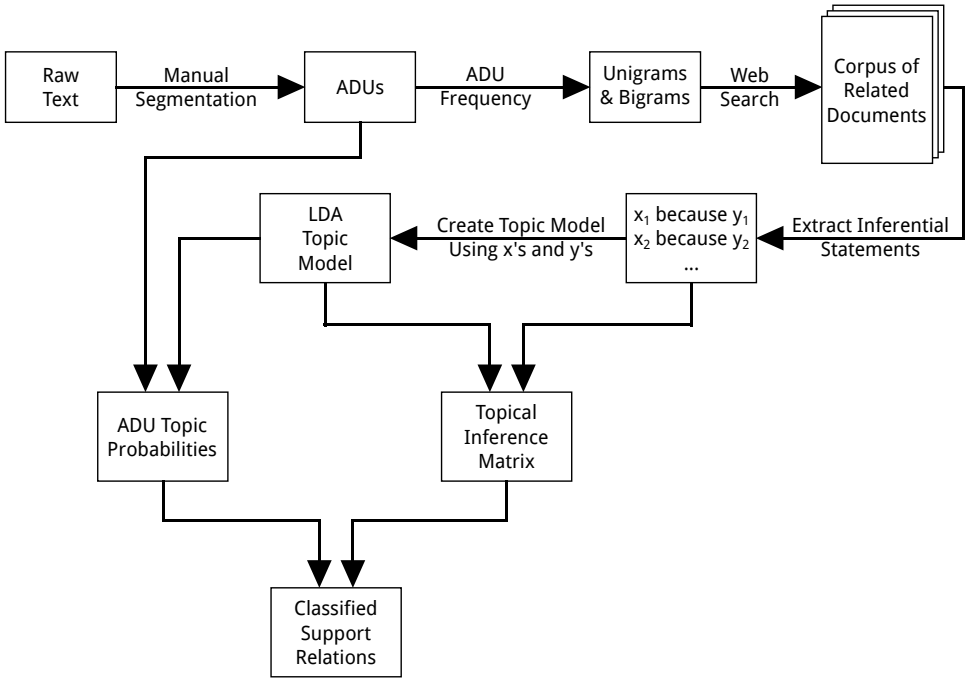

Figure 2: Overview of the Implementation Methodology for Creating Extended Corpus, Creating a Topical Inference Matrix and Classifying Support Relations

Unigram	Count
investment	39
banking	35
banks	28
money	27
problem	16
capitalism	13
culture	12
behaviour	12
rules	12
ethical	10

Table 1: Top ten unigrams by number of ADUs in which they appear

Bigram	Count
investment banks	18
investment banking	12
common good	5
immoral behaviour	3
free market	3
banking industry	3
wealth creation	3
redeemed capitalism	2
moral code	2
dutch bankers	2

Table 2: Top ten bigrams by number of ADUs in which they appear

duplicates, a total of 6,981 pages remained.

Although the pages identified in the previous step are high ranking search results for the terms identified, such pages commonly contain material unrelated to the topic, for example, advertisements and summaries of other articles. In order to extract those sections of the documents most likely to contain the body of an article, the Python *Beautiful Soup* library[3] was used to parse the HTML and extract consecutive paragraphs of text.

These paragraphs were then split into sentences, using the NLTK[4] tokeniser, and each of the resulting sentences searched for the presence of a discourse indicator. Previous work using discourse indicators to identify argumentative structure (Lawrence and Reed, 2015) has shown that, although not common enough to give a full representation of the structure, when present, discourse indicators give a very clear indication of the argumentative connection between two spans of text. As our aim is to extract only those sentences most likely to contain an inferential relationship, we first looked more closely at the relative performance of different indicators. Based on analysis of a separate *Moral Maze* episode, we identified those indicators showing the highest precision (the

[3] http://www.crummy.com/software/BeautifulSoup/

[4] http://www.nltk.org/

Indicator	Precision	Recall
therefore	0.95	0.0004
because	0.91	0.0031
consequently	0.82	0.0001
hence	0.76	0.0001
accordingly	0.74	0.0002
so	0.73	0.0005
after	0.69	0.0011
since	0.65	0.0008
then	0.58	0.0013
for	0.57	0.0006

Table 3: Top ten discourse indicators sorted by precision

precision and recall for the top ten indicators can be seen in Table 3). These results show that, when present, "therefore" and "because" give the highest indication of inference with a significant drop in accuracy for the remaining indicators. As such, we limited our generated corpus to only those sentences containing one of these two words.

Where the number of words either before or after the matching indicator was less than 5, the sentence was discarded. After carrying out this process, a total of 7,162 inferential sentences were identified (6,288 containing "because" and 874 containing "therefore"), giving a dataset of premise conclusion pairs, either *premise* ***therefore*** *conclusion* or *conclusion* ***because*** *premise*.

Whilst we do not have 100% precision for either of the discourse indicators used, the impact of this is mitigated by the way in which the resulting pairs are subsequently used. The use of the topic models described in the next section means that we neither need *all* of the inferential relations contained within our search results, or for *every* premise conclusion pair to be correctly labelled as such. The models which we produce may have a small amount of noise generated by false-positives, but these either comprise topics which are not then matched to elements from the original text, or add a small number of lower importance terms to a valid topic.

4.2 Creating the Topical Inference Matrix

To extract the topical nature of the premise conclusion pairs previously identified, a Latent Dirichlet allocation (LDA) topic model was created using the Python gensim library[5]. To produce this topic model, the sentences were first split where the indicator occurred, giving two documents for each sentence (one representing a premise, and the other, the conclusion). For our experiments, the model was created with forty topics using 20 passes over the supplied corpus.

From the probability distributions for each pair of conclusion (C) and premise (P), a topical inference matrix (T) was created, where the *i,j*th entry in the matrix corresponds to the product of probabilities that the premise has topic *i* and the conclusion topic *j*. For example, in the simplest case, if there is a probability of 1.0 that the premise has topic *m* and the conclusion topic *n*, then the matrix will contain 1.0 at *m,n* and zero for all other possible pairings. So, given topic distributions θ^C for the conclusion, and θ^P for the premise, T is defined thus:

$$t_{i,j} = \theta^P_i * \theta^C_j \tag{1}$$

To investigate the validity of our assumption that there would be a noticeable pattern in the relationships between topic and inference, we first created a combined topical inference matrix for each of the *because* relations identified, by summing all of the matrices resulting from these relations. We then looked at the entropy of this matrix calculated as the sum of the differences between each value in the matrix and the mean of all values. For the *because* matrix, the mean score was 3.67 and the total difference was 2275.58, giving an average difference of 1.42 for each item in the matrix from the mean value (with no relationship between topic and inference, this difference would be $\sim$0).

A corresponding matrix was then produced for the *therefore* relations, and the distance between the *because* and *therefore* matrices calculated. This calculation was performed by first scaling the values in each matrix to a value between zero and one, and then calculating the distance between the resulting matrices:

$$d(A, B) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{i,j} - b_{i,j})^2} \tag{2}$$

For identical matrices, this distance would be zero, for a pair of 40×40 matrices where all entries have maximal difference, the distance would be 40, and for a pair of 40×40 matrices where all entries have an average difference of 0.5 (in-

[5] https://radimrehurek.com/gensim/

dicating no correlation between the two), the distance would be 28.29. The distance between the *because* and *therefore* matrices was calculated as 18.32, suggesting a positive correlation between the two. We are not aware of any other technique that can be used to quantify the significance between such datasets: our analysis indicates merely that there is indeed some pattern beyond random chance linking the two concepts.

Finally, the *because* and *therefore* matrices were summed to give an overall topical inference matrix.

5 Experiments

In order to test our original hypotheses that the thematic commonalities present in the expression of human reasoning can be identified and used to help determine the structure of that reasoning, a number of experiments were carried out to explore the effectiveness of using this data to determine both the direction of inference between two ADUs that are known to have an inferential relationship, and the connectedness of pairs of arbitrary ADUs.

5.1 Using the Topical Inference Matrix to determine directionality

The manual analysis of our original text contained 128 premise conclusion pairs. As an initial experiment, we investigated how well the produced topical inference matrix could determine the direction of the inference between these pairs. This was achieved by creating a test set containing each pair (*a,b*) and its reverse (*b,a*).

Two alternative methods were tested to classify these pairs as being 'inferential' or 'non-inferential'. In each case, the topic probabilities for the ADUs were first inferred from the LDA model and a score determined as to whether there was an inferential relationship. For the first method, (MaxTopic), the score was calculated by taking the highest probability topic for each ADU and using these to look up the corresponding value in the overall topical inference matrix:

$$S_{MaxTopic} = t_{max(\theta^P),max(\theta^C)} \quad (3)$$

For the second method, (TopicDist), the values in the matrix were multiplied by the corresponding probabilities for each item in the pair and then summed to give an overall score.

$$S_{TopicDist} = \sum_{i=1}^{n}\sum_{j=1}^{n} t_{i,j} * \theta_i^P * \theta_j^C \quad (4)$$

For each of these two methods, the resulting scores were then compared against the mean of all values in the matrix (mean $= 3.15$), over which a pair would be classified as being 'inferential', and below which, 'non-inferential'.

Method	Precision	Recall	F_1-score
Random Baseline	0.5	0.5	0.5
MaxTopic	0.51	0.82	0.63
TopicDist	**0.57**	**0.83**	**0.67**

Table 4: Results for the MaxTopic and TopicDist methods to determine directionality of inferential connections compared to the random baseline

The results for directionality can be seen in Table 4. The results show an improvement over the random baseline for both methods, however the improvement in precision is low when just looking at the highest scoring topic. One reason for this is that a reasonable percentage of pairs (twenty-five out of one hundred and twenty-six) have the same highest scoring topic for both items (i.e. a conclusion is being supported by a premise that is closely related). When these same topic pairs are removed, the precision increases to 0.56, comparable to the results for the weighted topic distribution. The results for using the weighted topic distribution are better, and suggest that even in cases where the main topic is similar, there is enough of a difference in the secondary topics to determine the directionality of the pair.

5.2 Using the Topical Inference Matrix to determine connectedness

The second experiment performed looked at whether the produced topical inference matrix could determine inferential connections between arbitrary pairs of ADUs. For this task, a dataset was created containing the known 126 premise conclusion pairs and an equal number of random, unconnected ADUs. The same two methods of classifying these pairs as being 'inferential' or 'non-inferential' were used as in the first experiment, and the results can be seen in Table 5.

The results show that the precision is increased for classifying pairs as being connected over the previous results for directionality.

Method	Precision	Recall	F_1-score
Random Baseline	0.5	0.5	0.5
MaxTopic	0.58	0.79	0.67
TopicDist	**0.60**	**0.82**	**0.69**

Table 5: Results for the MaxTopic and TopicDist methods to determine connectedness of ADU pairs

5.3 Thresholding Topical Values

The experiments presented so far have looked at the likelihood that one topic supports another in terms of its score relative to all other scores in the matrix. However, it is possible that for some topics the scores will generally be higher. For example, if a large number of propositions have a high probability of corresponding to topic *n*, then all the values in column *n* of the matrix will be disproportionately high. To overcome any problems caused by this kind of topical skew, we took each column of the matrix and divided each value by the sum of values in that column. This resulting scaled matrix was then used to perform the same experiments as previously. The results for both experiments combined are shown in Table 6.

Method	Precision	Recall	F_1-score
Directionality			
Random Baseline	0.5	0.5	0.5
MaxTopic	0.61	0.77	0.68
TopicDist	**0.65**	**0.78**	**0.71**
Connectedness			
Random Baseline	0.5	0.5	0.5
MaxTopic	0.59	0.75	0.66
TopicDist	**0.64**	**0.83**	**0.72**

Table 6: Results for the MaxTopic and TopicDist methods to determine connectedness and directionality using a thresholded inference matrix

In all cases, we can see that the precision is slightly improved, though (with the exception of the TopicDist results for connectedness) this is at the expense of recall.

6 Discussion

The results we have presented show in all cases that there is some correlation identified between the topics that a pair of ADUs have, and the nature of their potential inferential relationship. By looking at the topics of each item in the pair, we have been able to determine both connectivity and directionality of inference. Overall, the results are better for identifying connectedness than directionality, predominantly resulting from higher similarity in topics for which the ADUs are connected (in a significant percentage of cases the maximum probability topic was the same).

Currently, the identification of relationships is limited to inferential relationships, and one area of development would be to extend this by examining those discourse indicators which show a conflict relationship. Additionally, no account is taken of the polarity or sentiment of the ADUs. Where we have a conclusion, 'C', and a premise, 'P', then there would be a high topical similarity between P and 'not P', and as such, an inference relationship would be assigned between them. This problem could be overcome by applying sentiment classification to the ADUs as a preliminary step, and where there is negation of one item in the pair, replacing an inference relationship with conflict. Expanding the scope of this technique to give a fuller indication of relations will be carried out in future work.

Although we focus on identifying patterns of inference within a single debate, there is nothing intrinsic to the approach that makes it a better fit for this domain than any other. The automatic determination of the domain being discussed requires only the original text, and from this we are able to build a dataset specific to that domain which, due to the reliability of discourse indicators, contains domain specific pairs that we can say with high confidence have an inferential relationship.

7 Conclusion

This work has demonstrated how by automatically creating large, high-confidence datasets of inferential pairs related to a specific topic, we can closely mirror one of the ways in which humans understand the complex interactions between the individual propositions expressed in a debate.

The approach presented is effective in tackling the challenging high-level pragmatic task of identifying both connectedness and directionality between argumentative discourse units, with results 22 points above baseline.

This outcome represents strong performance for this level of task (cf., for example, (Feng and Hirst, 2011; Peldszus, 2014)), giving results comparable to those of (Palau and Moens, 2009), where each Argument sentence was classified as either premise or conclusion with F_1-scores of 0.68 for classification as premise and 0.74 for conclusion. Furthermore, where existing approaches are often

constrained in their generality by a lack of appropriately annotated, domain-specific, data, the same requirement does not apply in this case.

The results show a clear link between the words used to express an argument and its underlying structure, and strongly support the intuition that understanding the structure of an argument requires not only consideration of the text itself, but contextual knowledge and understanding of the broader issues. We see this work as a key component in a larger ensemble approach (Lawrence and Reed, 2015), mirroring the complex process followed by a human annotator whereby general domain knowledge, understanding of linguistic cues and familiarity with common patterns of reasoning are combined to understand the arguments being made.

Acknowledgements

This work was funded in part by EPSRC in the UK under grant EP/N014871/1.

References

Khalid Al-Khatib, Henning Wachsmuth, Hagen Matthias Stein, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of NAACL-HLT*. pages 1395–1404.

Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the argument web. *Communications of the ACM* 56(10):66–73.

David M. Blei, Aandrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*. pages 917–924.

Carlos Chesñevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, et al. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* 21(04):293–316.

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 (EMNLP)*. pages 2236–2242.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 987–996.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 2127–2137.

Hospice Houngbo and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 19–23.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: An argument analysis interface. In S. Parsons, N. Oren, C. Reed, and F. Cerutti, editors, *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*. IOS Press, Pitlochry, pages 463–464.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. AIFdb: Infrastructure for the argument web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*. pages 515–516.

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

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 79–87.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pages 343–351.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2:627–666.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2003. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03)*. IEEE, pages 179–186.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 20–28.

Raquel M. Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*. ACM, pages 98–107.

Vincenzo Pallotta and Rodolfo Delmonte. 2011. Automatic argumentative analysis for interaction mining. *Argument & Computation* 2(2-3):77–106.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 108–112.

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 88–97.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7(1):1–31.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC-2008)*.

Patrick Saint-Dizier. 2012. Processing natural language arguments with the <TextCoop> platform. *Argument & Computation* 3(1):49–82.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 46–56.

Maria Paz G. Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*. pages 23–34.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC)*. pages 812–817.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering* 18(4):437–490.

Adam. Wyner, Jodi. Schneider, Katie. Atkinson, and Trevor. Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*. pages 43–50.