Language Resources for Studying Argument

Chris Reed[1], Raquel Mochales Palau[2], Glenn Rowe[1], Marie-Francine Moens[2]

[1]School of Computing, University of Dundee, Dundee DD1 4HN, UK
[2]Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3000 Leuven, Belgium
[chris | growe]@computing.dundee.ac.uk, [raquel.mochales | sien.moens]@cs.kuleuven.be

Abstract

This paper describes the development of a written corpus of argumentative reasoning. Arguments in the corpus have been analysed using state of the art techniques from argumentation theory and have been marked up using an open, reusable markup language. A number of the key challenges enountered during the process are explored, and preliminary observations about features such as inter-coder reliability and corpus statistics are discussed. In addition, several examples are offered of how this kind of language resource can be used in linguistic, computational and philosophical research, and in particular, how the corpus has been used to initiate a programme investigating the automatic detection of argumentative srtucture.

1. Introduction

Argumentation theory as a discipline is focused on understanding the ways in which humans express their reasoning, articulate disagreement, and reach consensus (van Eemeren et al., 1996). The field has empirical and normative branches, and covers both monological and dialogical argument. Though based in philosophy, it abuts onto cognitive science, linguistics and communication theory. One can find arguments in newspapers (in editorial comment and letters in particular), political speeches (e.g. during hustings), legislative texts (e.g. arguments sustaining a certain norm), in case law (where analogical argument is particularly prevalent), opinion blogs and discussions boards (e.g. personal arguments attacking or defending the views of others), and doctrinal texts (e.g. arguments concerning the interpretation of a religious principle).

Over the past ten years or so, argumentation theory has increasingly been used by researchers in artificial intelligence to develop models and systems of defeasible reasoning, natural language processing, inter-agent communication and more (Reed and Grasso, 2007). With this explosion in computational models of argumentation, has come the need for structured representations of arguments, and the tools to analyse, manipulate and transform those representations. As these tools themselves have started to mature, they in turn have found applications back in philosophy and linguistics. This has triggered an increase in empirical linguistic study of argumentation with approaches and resources familiar to the linguists, such as the recent appearance of argument corpora.

2. Argument Corpora

The aim of current argument corpora research is twofold. First, by increasing coherence between the resources (e.g. in how they are represented and accessed), to develop a foundation that will support a wide range of small-scale individual argumentation related projects. Second, by linking and making possible resource reuse, to also enable multi-site, international, directed efforts at synthesising large-scale corpora, that can do for argumentation analysis (and its computational applications) what resources like the Penn Treebank have done for natural language processing.

One might think it could be possible to use general text corpora, i.e. resources not designed specifically for argument research, to study argumentation. However, using general corpus would require time-consuming and expensive manual search over very large amounts of texts to identify the argumentative section Though automatic text scanning could reduce the costs, such methods are currently too unreliable, and indeed argument corpora are being used to develop exactly these sorts of methods. To date, therefore, compilation of dedicated argument resources has been the only option.

The University of Dundee, aiming to develop both a reliable resource for researchers working on argumentation and a test case for future development of similar corpora, has developed the AraucariaDB corpus. This corpus represents, to the best of the team's knowledge, the first resource of its kind (though textual arguments are occasionally collected, as in, for example, David Hitchcock's set of student arguments, and the University of Durham's Free Britain Corpus, these examples do not include argument analysis or reconstruction).

2.1. Corpus Development

Work on the AraucariaDB corpus commenced in 2003, comprising a set of argumentative examples extracted from diverse sources (Table 1) and different regions, including India, Japan, South Africa, UK, Australia and the US, allowing a wide range of different argumentative styles. All source material is in English. Furthermore, all examples are available in their original form online, in most cases with an implication that they are available in perpetuity.

For each example, a time-based sub-sample mechanism was adopted. The collection of material drew about five items from each of the twenty sources over a period of several weeks, and each item was then subjected to detailed manual analysis. In order to build machine readable analyses of argument structure, original texts were extracted and marked up using Araucaria, a software tool that allows rapid graphical analysis of argument structure (Reed and Rowe, 2004). Propositional atoms can be selected from

Type	Source
Newspaper editorials	The Age, Talking Point (BBC Online), Have Your Say (BBC Online), Outlook (India), The Japan Times, Indian Express, The Independent, New York Times, Mail, Guardian Online (South Africa),The Telegraph, The Washington Post
Parliamentary records	High Court of England Wales, UK House of Lords Judgements, Indian Parliamentary Debates, UK House of Parliament Debates
Judicial summaries	United States Supreme Court, US Congressional Record
Discussion boards	Research Ministry Discussion Board, National Public Radio Discussion Board, Human Rights Watch, Christian Apologetics

Table 1: Sources

the text and dragged onto the drawing pane, and inferential relationships between them are created by dragging and dropping links between them. The result is stored in an XML-based format, the Argument Markup Language (AML). (See Figure 1 for an example of the diagramming notation and its rendering in AML).
The corpus was extended in 2004 with further analytical ontologies based on differing argumentation scheme sets. By the completion of the 2004 phase, over 700 analyses were available representing almost 4,000 atomic propositions, 1,500 reconstructed premises, and a total of 80,000 words. Of course, 80,000 words is a rather small corpus by the standards of corpora in general, but argument analysis is a highly labour intensive task. For example, 5,000 analysed argument components represent almost 12 man-months of work.
Both corpora versions, 2003 and 2004, were analysed according to theories of argument structure, paying special attention to each argument role and its components. In concrete we focused on a theory described by Walton (1996) based on argumentative scheme sets. Walton describes argumentation schemes as linguistic forms expressing stereotypical patterns of reasoning that form the 'glue' of interpersonal rationality. These schemes are becoming increasingly important in both argumentation theory and computer science (see, e.g., (Atkinson et al., 2006), (Reed and Walton, 2005), (Walton, 2005)). Figure 2 shows a diagrammatic rendering of a sample argumentative analysis performed using Araucaria, showing several argumentation schemes.

2.2. Development Issues

During the corpus development a number of key questions that frame challenges for argument corpora were raised. The ramifications of these challenges are here discussed, and some potential avenues towards tackling them are explored.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ARG SYSTEM "argument.dtd">
<ARG>
  <?Araucaria UTF-8?>
  <TEXT>Here are some bear tracks in
  the snow. Therefore, a bear passed
  this way. </TEXT>
  <EDATA>
    <AUTHOR>null</AUTHOR>
    <DATE>2003-05-09</DATE>
    <SOURCE />
    <COMMENTS />
  </EDATA>
  <AU>
    <PROP identifier="B" missing="no">
      <PROPTEXT offset="50">a bear
      passed this way</PROPTEXT>
      <INSCHEME scheme=
      " Argument From Sign "
      schid="0" />
    </PROP>
    <CA>
      <AU>
        <PROP identifier="A"
        missing="no">
          <PROPTEXT offset="0">Here are
          some bear tracks in the snow
          </PROPTEXT>
          <INSCHEME scheme=
          " Argument From Sign "
          schid="0" />
        </PROP>
      </AU>
    </CA>
  </AU>
</ARG>
```

Figure 1: The AML and diagrammatic forms of a sample file distributed with Araucaria.

2.2.1. Inter-coder reliability

The annotation of argumentation is clearly a subjective task. Therefore, as in other subjective tasks, the main worry was (and to an extent, continues to be) that inter-coder reliability would simply be too low to give enough confidence in analytical results. Of course, this may simply be a reflec-

The Bali bombings are the work of an international terrorist group, not just local Islamic radicals. For example, the bomb used in the nightclub attack was reportedly made from a military plastic explostive similar to the one used in the attack on the USS Cole in Yemen two years ago.

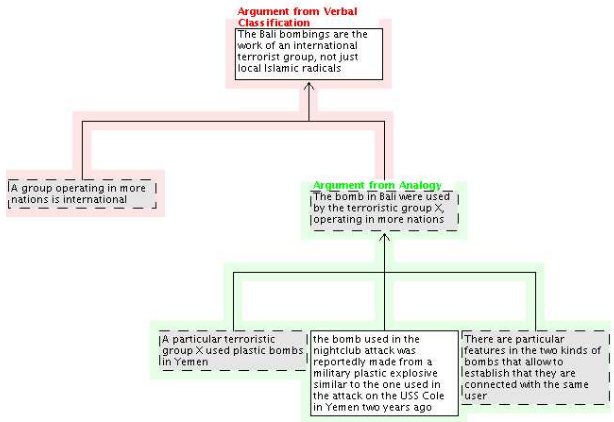

Figure 2: A sample text from the corpus and one of its analyses in the corpus (*The Japan Times*, Op-Ed, "Most Crucial Lesson from Bali", 18 October 2002).

tion of the inadequacy of the analytical techniques themselves. However, intuitions from conducting analyses and from teaching others how to do so, seemed to support that more or less "right" analyses do exist for most texts. Moreover, most of these analysis are also "unique" or where there are differences of analytical judgement, there are usually a relatively small number of alternative valid assessments.

The large-scale analysis of argumentative texts had no precedent that we were aware of, so in discussion with the two analysts, it was agreed that each would work to their style, within the general constraints of Walton's argumentation theory described above. The first analyst invested a great deal of time and effort in each analysis yielding high precision; the other proceeded more quickly, with a slightly higher error rate. (The "precision" and "error rate" here are not quantified: they are the result of subjective assessment of quality based on subsampling and review by other members of the team in informal discussions). The second analyst was also inclined to perform somewhat more reconstruction (i.e. to more often add in material that was not explicitly stated in the original text). Such reconstruction is entirely licit – it is a common part of argument analysis (van Eemeren et al., 1996) – but as with many other analytical linguistic endeavours, the degree of reconstruction is variable.

There are several interesting observations in analysing inter-coder reliability, but one that is particularly intriguing is that the first analyst identified enthymemes somewhat more frequently than the second. At first blush this is in contradiction with the fact that it was the second analyst who introduced slightly more material. Specifically, the first analyst introduced 2.45 enthymemes on average in each analysis; the second introduced 2.53. However, the second also made slightly more detailed analyses (identifying 6.18 propositions on average in each analysis as opposed to 5.43 for the first), so the occurrence of enthymemes over a given number of propositions was actually higher for the first analyst. Though the difference (from one enthymeme every 2.2 propositions to one every 2.4) may seem slight, it is useful to put this into context. A ratio of 1 in 2.0 is a very high value, meaning that for every pair of propositions an enthymeme has been reconstructed. It might be expected that the scale is logarithmic such that a ratio of 1 in 4.0 would mean that half of all proposition-pairs are so reconstructed. In fact, however, the number of inferences is not half the number of propositions: many arguments in the corpus, for example, have three explicit propositions arranged in a serial chain. In such a case, there are three propositions linked by two inferences. To interpret the frequency of enthymemes, therefore, one has to bear in mind that there are on average around 1.4 propositions to each inference. So 1 in 1.4 would mean reconstructing every inference as an enthymeme, and 1 in 2.8 reconstructing half of all inferences as enthymemes. Bearing this scale in mind, the difference between the two coders equates to something like a difference of opinion on one in every 14 inferences.

The experience with the two coders above suggests that the differences between annotators might be more slight than had been feared. However, from a methodological point of view, it will be important to assess, track and improve intercoder reliability. At the moment, we have little experience of how coders differ when working within a given argumentative framework, nor any experience of how argumentation theoretic analysis frameworks differ in terms of their ability to explain data, and their ability to successfully guide coders to consensus analyses. Perhaps results from pedagogic investigations, such as (Hitchcock, 2002), might be used to design ways of increasing harmonisation between coders, and perhaps software tools that impose greater restrictions might help too. In fact, argumentation schemes may, through their critical questions, provide one way of imposing such restrictions semi-automatically. Therefore, we believe such schemes form a key step in the development of a generic coding scheme.

2.2.2. Coding & Reusability

The development of a common coding scheme is a key requirement in enhancing reusability (much as TEI has demonstrated in other linguistic domains). The first step in this direction was the emergence of Araucaria's AML language for marking up resources (Reed and Rowe, 2004). This language was open and free, and designed to be reused. However, by the time the Araucaria software became a commonplace, AML was (exceedingly) long in the tooth with numerous restrictions and limitations that were overconstraining, such as the exclusion of divergent arguments or the inability to handle defeasibility adequately.

Other representation formats are either closed, and therefore difficult to re-use, or else intimately tied to a specific software application. This gap in the market, combined with the need for exchange of arguments between different software applications and different research groups led to the development of the Argument Interchange Format (Chesñevar et al., 2006). Software tools that use the AIF (including an updated version of Araucaria) are now in the pipeline all over the world. It is expected that the first version may have some few deficiencies, but the hope is that AIF will provide an extensible framework in which all those who work with argument including those who build and manipulate argument corpora might have something to gain through its use.

2.2.3. Cost

The third and final challenge is one of cost. Getting good analyses can be very time consuming (of the order of hours for a several hundred word text). This represents a practical, mundane - but very real - barrier to large corpus creation. Software systems may also, possibly, help here too. At the moment, argumentation corpora hold huge potential but without some investigations it is difficult to be sure of exactly what might be gained. Preliminary and typically small-scale investigations such as the one described here are starting to sketch out the space. But until these research tasks can be brought together in such a way as to evidence larger project proposals with larger scale funding, the human labour costs simply prohibit the construction of large, manually analysed corpora. One interesting avenue is to build tools not for the analysis of argument, but rather for its intuitive and straightforward construction by non-expert users. If these tools provide simple easy-to-use interfaces, whilst preserving rich argumentation-theoretic structures under the hood then it becomes possible to harness the enormous manpower of internet users: allowing users to construct arguments in a form that is pre-analysed could make available to argumentation corpus researchers what has long been available to unanalysed text corpora, namely, the sheer size of the internet. The key, of course, is to make sure that these software systems are both sufficiently easy to use and sufficiently beneficial to users that they are in fact adopted. Initial steps towards tools that fit that bill are being made (Kirschner et al., 2003), (Rahwan et al., 2007).

2.3. Analysis

Reed (2005) discusses a range of the features, characteristics and results drawn from the 2003-corpus analysis, from which we summarise a few of the more important here. The first, and most prominent, feature of the dataset is the preeminence of normative argument, and specifically, of the two schemes in the Katzav & Reed (2004) taxonomy, Argument from the Constitution of Positive Normative Facts and its counterpart, Argument from the Constitution of Negative Normative Facts. It is interesting that normative arguments with a clearly positive conclusion are much more common that those with a clearly negative conclusion by a factor of around two and one half. This may be as a result of a rhetorical rule based at least in part in the social psychology of message adoption (McGuire, 1968) - positive conclusions are more likely to be accepted than their negatively phrased counterparts. Some domains, however, show distinct identities in terms of the argumentation schemes that are employed, and this is a second observation.

A good example is the scheme Argument from Implication, which explicitly builds a deductive structure. Although not entirely uncommon, the overwhelming majority occur in newspaper and magazine editorials. One possible explanation for the disproportionately high frequency of the scheme in popular press editorials concerns expectation and appearance. Editorials are supposed to be strongly argumentative, with a clear standpoint in the pragma-dialectical sense (van Eemeren and Grootendorst, 1992). One of the ways of conveying such clarity and of developing a strong, characteristic argumentative flavour, is to use relationships between discourse components which themselves have clear argumentational roles. Argument from Implication fits this bill admirably. Further support for this contention is offered by the fact that Argument from Implication is often associated with strong clue words such as *therefore*, *because*, and *as a result* which signpost an argument, making its structure clearer to the reader – and thereby also making clearer the fact that it is an argument. Of course, this role for clue words is well known both in (computational) linguistics (Knott, 1996) and in argumentation theory (Snoeck Henkemans, 2003) - in the latter, it is often used as a mechanism for helping students learn first to identify and then to analyse instances of argumentation (see, e.g. a textbook such as (Wilson, 1986)[pp17-23]). Key words and other surface features available from corpus collection can be used to train classifiers for argument detection. This works particularly well in specific domains, as we explore briefly in the next section.

3. Applications

The corpora presented in this paper opens new research areas as well as new techniques to achieve older objectives of Natural Language Processing. It will be a useful tool to extend different research aspects on argumentation and discourse, such as the following:

- *An empirical evaluation of (Walton, 1996) argumentation theories.* The analysis of our compiled documents can be used to detect the main schemes on real written argumentation, the main sources by scheme or the prefered schemes depending on the target audience.

- *Improvement of discoursive or rhetorical analysis.* The main syntactical and semantical structures in the argumentation process can be discussed based on our example corpora of real written argumentation.

- *Learning critical thinking.* Our corpora facilitates the teaching of critical thinking, allowing the students to easily learn the main characteristics and structures of this argumentative process by example. Araucaria and its corpus are in use in hundreds of university teaching environments worldwide.

A promising new research area, offered by our corpora, is the automatic detection of arguments in discourses. Automatic argumentation detection is an important task in Case Base Reasoning, text summarization, meeting tracking and information visualization with a wide range of applications. Because of the complex structure of argumentative

Feature	Description
Unigrams, Bigrams and Tigrams	Each word in the sentence, each pair of successive words and each three successive words.
Adverbs and Verbs	Only the main verbs (excluding “to be”, “to do” and “to have”) were considered to study.
Modal Auxiliaries	To indicate the level of necessity or conditionality of a sentence.
Word Couples	All possible combinations of two words in the sentence were considered. Only cleaned couples (not containing “to be”, “to do”, determiners, such as “a” or “the”, proper nouns or symbols).
Punctuation	Different sequences of punctuation marks.
Textual Statistics	Sentence length, average word length, punctuation frequency, etc.
Key words	List of 286 words, such as “but” or “consequently”.
Parse complexity	Depth of the parse tree of each sentence, number of subclauses, etc.

Table 2: Trained features on (Moens et al., 2007)

discourse and the lack of resources it has been left nearly unstudied till the moment.

Automatic argumentation detection can be divided in two main tasks: (a) the detection of an argument, its boundaries and its relations with other text sections, and (b) the detection and classification of the different components that make up the argument, i.e. the recognition of premises and conclusions. Both tasks require an extensive use of argumentative corpora. However, while task (a) demands full argumentative text analysis, task (b) is based on the analysis of isolated arguments.

The 2003 corpus has been used as the initial resource to solve task (b) in (Moens et al., 2007), where automatic detection of argumentative and non-argumentative sentences is studied. The main objective of this work was to detect if a sentence contained an argumentative fragment, i.e. a premise or a conclusion. This study was further extended in (Mochales Palau and Moens, 2007) where there was a more fine-grained detection of argumentative fragments, distinguishing between premises and conclusions. However, the need for contextual information in this work required the use of a different corpus than the 2003 corpus presented in this paper. Both studies threated argumentation detection as a classification problems where different state-of-the-art classification algorithms, e.g. a naive bayes classifier, were studied using a manually annotated corpus, e.g. the 2003 corpus, and different feature vectors.

The trained features (Table 2) used in these tests, even though only an initial assessment for the identification of arguments in single sentences, achieved a 74% accuracy and brought up some interesting remarks on argumentative discourse analysis. For example, the poor discrimination rate achieved with the study of keywords denoted a high ambiguity of their use in both statements and arguments. Also the low rates achieved with modal auxilaries reflected a tendency in written discourse to use similar syntactical and structural styles when presenting conditional facts and arguments.

4. Conclusion

We have presented the development of a language resource for argumentation analysis, together with some experiences with our initial pilot data collection, which raised a number of key questions that frame challenges for argument corpora in general. To the best of the team’s knowledge, this corpus represents the first resource of its kind, and is currently being utilised by software systems in both teaching and research contexts. Furthermore, the increase of studies on defeasible reasoning, written argumentation analysis and inter-agent communication open new applications for this kind of resources.

One retort to the methodological challenges summarised here and discussed in more detail in (Reed, 2005) is simply to see them as being a result of the goals of this or any project. So, for example, the fact that there are multiple sets of argumentation schemes, necessitating a corpus that can admit analyses based on different such sets, is simply a result of the fact that this project is interested in looking at argumentation schemes – it is not a general problem at all. Similarly, defining the source material, defining the collection regime, identifying arguments, and analysing those arguments are, it might be claimed, all tasks that are dependent on the goals of a research project. Such a line of reasoning is specious. One of the key jobs that a corpus can play is in providing a foundation for multiple projects. The most successful corpora are not just used by hundreds of research projects, but they enter the shared cultural backdrop for researchers in dozens of academic fields: the Brown corpus, the BNC and the Penn Tree Bank are perfect examples of this. Such corpora have been assembled in such a way that ever more new hypotheses can be tested, ideas explored, and projects constructed. To say that argument corpora can only be formed once the goals of a specific project are known is to permanently restrict the scope of what they can support. What we hope to have demonstrated here is that a single corpus is now starting to be used in this broader way, opening up many new possibilities for the development and widespread use of argument language resources.

5. Acknowledgements

We would like to acknowledge the support of The Leverhulme Trust in the UK which contributed towards some of the initial corpus development, and we would like to acknowledge the work of the coders, Fabrizio Macagno and

Joel Katzav, and the work on the web interface built by Louise McIver.

6. References

K. Atkinson, T. Bench-Capon, and P. McBurney. 2006. Computational representation of practical argument. *Synthese*, 152(2):157–206.

C. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott. 2006. Towards an argument interchange format. *Knowledge Engineering Review*, 21(4):293–316.

D. Hitchcock. 2002. Toulmin's warrants. In *Proceedings of the 5th International Conference of Argumentation (ISSA 2002)*. SicSat.

J. Katzav and C. Reed. 2004. On argumentation schemes and the natural classification of argument. *Argumentation*, 18(4):239–259.

P. Kirschner, S. Buckingham Shum, and C. Carr. 2003. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer.

A. Knott. 1996. *A Data Driven Methodolgy for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

W.J. McGuire. 1968. Attitudes and attitude change. In G. Lindzey and E. Aronson, editors, *The Handbook of Social Psychology*, pages 136–314.

R. Mochales Palau and M-F. Moens. 2007. Study on sentence relations in the automatic detection of argumentation in legal cases. In *Proceedings of JURIX 2007: The 20th Anniversary International Conference on Legal Knowledge and Information Systems*.

M.F. Moens, E. Boiy, R.M. Palau, and C. Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the International Conference on AI & Law (ICAIL-2007)*.

I. Rahwan, F. Zablith, and C. Reed. 2007. Laying the foundations for a world wide argument web. *Artificial Intelligence*, 171:897–921.

C. Reed and F. Grasso. 2007. Recent advances in computational models of argument. *International Journal of Intelligent Systems*, 22(1):1–15.

C. Reed and G.W.A. Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 14(3-4):961–980.

C. Reed and D. Walton. 2005. Towards a formal and implemented model of argumentation schemes in agent communication. *Autonomous Agents and Multi-Agent Systems*, 11(2):173–188.

C. Reed. 2005. Preliminary results from an argument corpus. In *Proceedings of the IX Symposium on Social Communication*, pages 576–580.

A.F. Snoeck Henkemans. 2003. Indicators of analogy argumentation. In *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation (ISSA-2002)*, pages 969–973. SicSat.

F.H. van Eemeren and R. Grootendorst. 1992. *Argumentation, Communication and Fallacies*. Lawrence Erlbaum Associates.

F.H. van Eemeren, R. Grootendorst, and F.A. Snoek Henkemans. 1996. *Fundamentals of Argumentation Theory*. LEA.

D.N. Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates.

D. Walton. 2005. *Argumentation methods for artificial intelligence in law*. Springer.

B. Wilson. 1986. *The Anatomy of Argument*. University Press of America.