The CASS Technique for Evaluating the Performance of Argument Mining

Rory Duthie[1], **John Lawrence**[1], **Katarzyna Budzynska**[1,2], **and Chris Reed**[1]

[1]Centre for Argument Technology, University of Dundee, Scotland
[2]Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland

Abstract

Argument mining integrates many distinct computational linguistics tasks, and as a result, reporting agreement between annotators or between automated output and gold standard is particularly challenging. More worrying for the field, agreement and performance are also reported in a wide variety of different ways, making comparison between approaches difficult. To solve this problem, we propose the CASS technique for combining metrics covering different parts of the argument mining task. CASS delivers a justified method of integrating results yielding confusion matrices from which CASS-κ and CASS-$F1$ scores can be calculated.

1 Introduction

To calculate the agreement, or similarity, between two different argumentative structures is an important and commonly occurring task in argument mining. For example, measures of similarity are required to determine the efficacy of annotation guidelines via inter-annotator agreement, and to compare test analyses against a gold standard, whether these test analyses are produced by students, or automated argument mining techniques (*cf.* (Moens, 2013; Peldszus and Stede, 2013)).

To find the the similarity of automatic and manually segmented texts and what impact these segments have on agreement between annotations for an overall argument structure, is a complex task. Similar to these problems is the task of evaluating the argumentative structure of annotations using pre-segmented text. Despite the relative ease of manually analysing these situations, arguments with long relations can easily make this task complex.

Commonly to find the agreement of manual annotators or the effectiveness of an automatic solution, two scores are given, Cohen's kappa (Cohen, 1960), which takes into account the observed agreement between two annotators and the chance agreement, giving an overall kappa value for agreement, and $F1$ score (Rijsbergen, 1979), which is the harmonic mean of the precision and recall of an algorithm. The way in which these scores are utilised can over penalise differences in argumentative structures. In particular, if used incorrectly, Cohen's kappa can penalise doubly (penalise for segmentation and penalise segmentation in argumentative structures) if not split into separate tasks or penalise too harshly when annotations have only slight differences, again if the calculation is not split by argumentative structure. When using the $F1$ score the same problems arise without split calculations.

To combat these issues this paper introduces two advances: first, the definition of an overall score, the Combined Argument Similarity Score (CASS), which incorporates a separate segmentation score, propositional content relation score and dialogical content relation score; and second, the deployment of an automatic system of comparative statistics for calculating the agreement between annotations over the two steps needed to ultimately perform argument mining: manual annotations compared with manual annotations (corpora compared with corpora) and automatic annotations evaluated against a gold standard (automatically created argument structures compared with a manually annotated corpus).

2 Related Work

Creating the CASS technique and an automatic system to calculate it, is based on theories established in linguistics and computational linguistics.

Proceedings of the 3rd Workshop on Argument Mining, pages 40–49,
Berlin, Germany, August 7-12, 2016.

In (Afantenos et al., 2012), a discourse graph is considered and split into discourse units and relations, to calculate agreement using $F1$ score. This gives what is described as a "brutal estimation" which gives an underestimation of the agreement. To combat this it is suggested that reasoning over the structures is needed.

In (Artstein and Poesio, 2008) a survey is given of agreement values in computational linguistics. Different measures of the statistics both Cohen's kappa and Krippendorff's alpha (Krippendorff, 2007) along with other variations are considered for different tasks. On the task of segmentation, it is noted that kappa in any form does not account for near misses (where a boundary missed by a word or two words) and that instead other measures (see Section 4) should be considered. On the topic of relations and discourse entities, again kappa in its various forms and alpha are considered. For both relations and discourse entities the kappa score is low overall because partial agreement is not considered. Instead the idea of a partial agreement coefficient is introduced as being applicable.

In (Habernal and Gurevych, 2016), Krippendorff's unitized alpha (α_U) is proposed as an evaluation method, to take into account both labels and boundaries of segments by reducing the task to a token level. The α_U is calculated over a continuous series of documents removing the need for averaging on a document level, but is dependent on the ordering of documents where the error rate of ordering is low.

Finally, in (Kirschner et al., 2015) methods for calculating inter-annotator agreement are specified: adapted percentage agreement (APA), weighted average and a graph based technique (see also Section 3.2).

APA takes the total number of agreed annotations and divides it by the total number of annotations, on a sentence level of argument but not corrected for chance. Chance is taken into account, when performing the weighted average. A weight is provided for the distance between related propositions when the distance is not greater than six. Meaning any relation with a distance greater than six is discounted. This is justified with only 5% of relations having a distance greater than two. Chance is accounted for by using this weighted average for multi-annotator kappa and $F1$ score. Finally, a graph based approach is defined, where the distance between nodes is taken for each annotation with each node distance as a fraction. The distance is added, then multiplied by the overall number of edges giving a normalised score for both annotations, not considering the direction or types of relations or any unconnected propositions. The harmonic mean is then taken to provide the agreement between the annotations.

Results are also provided when considering relation types for weighted average and nodes with distance less than six for inter-annotator agreement on propositional content nodes for a pre-segmented text.

If we consider the papers submitted to the 2nd workshop on argumentation mining, we can see there is an inconsistency in the area when calculating inter-annotator agreement and overall argument mining results. To calculate the agreement between annotators, three papers used Cohen's kappa (*cf.* (Bilu et al., 2015; Carstens and Toni, 2015; Sobhani et al., 2015)), three papers used inter-annotator agreement as a percentage (*cf.* (Green, 2015; Nguyen and Litman, 2015; Kiesel et al., 2015)), two used precision and recall (*cf.* (Sardianos et al., 2015; Oraby et al., 2015)) and three others used different methods (*cf.* (Kirschner et al., 2015; Yanase et al., 2015; Reisert et al., 2015)). To calculate the results of argument mining, four papers used accuracy (*cf.* (Bilu et al., 2015; Kiesel et al., 2015; Nguyen and Litman, 2015; Yanase et al., 2015)) and five papers used precision, recall and $F1$ score (*cf.* (Lawrence and Reed, 2015; Sobhani et al., 2015; Park et al., 2015; Nguyen and Litman, 2015; Peldszus and Stede, 2015)) with one paper using a macro-averaged $F1$. What is required in the area of argument mining is a coherent model to give results for both annotator agreement but also the results of argument mining.

In the area of text summarization, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) was created exactly for the purpose of having a coherent measure to allow systems in the Document Understanding Conference (DUC) to be evaluated. In creating the CASS technique we aim to emulate ROUGE, and provide consistency in the area of argument mining.

3 Foundation

3.1 Representing Argument

Arguments in argument mining can be represented in many forms which is particularly important for

the development of the CASS technique, as this score must be applicable to different ways of representing argument.

Scheme for Argumentative Microtexts. In (Peldszus and Stede, 2013) an annotation scheme was defined which was incorporated in a corpus of 122 argumentative microtexts (Peldszus and Stede, 2015). In this annotation scheme an argument is defined as a non-empty premise, that is a premise which holds some form of relation which supports a conclusion. Graphically this is represented by proposition nodes and support relations, with support relations represented as an arrow between the node and its conclusion.

The scheme defined builds on and extends the work of (Freeman, 2011). Support relations are defined in the most basic way, as an argument in the form of premise and conclusion. This accompanied by attack relations where rebutting is defined for when an argument is attacked directly and undercutting when a premise is attacked. Counter attacks then allow rebuttals of an attack support, the undercutting of an attack support and a counter consideration argument. Each microtext is pre-segmented to avoid bias from annotators segmenting text in their own style, with rules defined in the scheme which allow annotators to change the segmentation.

Internet Argument Corpus (IAC). Argument data is also represented use quote-response pairs (QR pairs) in the IAC (Walker et al., 2012). The IAC provides 390,704 individual posts automatically extracted from an Internet forum. Each post is related to a response which is provided through a tree structure of all the posts on the forum.

QR pairs work with a pre-defined segmentation which can allow annotators to identify relations between a quote (post) and a response. Relations can be on a number of levels with the most basic of these, agree and disagree, to the more complex, sarcasm where an annotator decides if a response is of sarcastic manner using their own intuition where a formal definition or annotation is near impossible without being present during the vocalisation of the point.

Argument Interchange Format (AIF). Argument data can also be represented according to the AIF (Chesñevar et al., 2006) implemented in the AIFdb[1] database (Lawrence et al., 2012). The AIF was developed as a means of describing argument networks that would provide a flexible, yet semantically rich, specification of argumentation structures. Central to the AIF core ontology are two types of nodes: Information- (I-) nodes (propositional contents) and Scheme (S-) nodes (relations between contents). I-nodes represent propositional information contained in an argument, such as a conclusion, premise etc. A subset of I-nodes refers to propositional reports specifically about discourse events: these are L-nodes (locutions).

S-nodes capture the application of *schemes* of three categories: argumentative, illocutionary and dialogical. Amongst argumentative patterns there are inferences or reasoning (RA-nodes), conflict (CA-nodes) and rephrase (MA-nodes). Dialogical transitions (TA-nodes) are schemes of interaction or protocol of a given dialogue game which determine possible relations between locutions. Illocutionary schemes are patterns of communicative intentions which speakers use to introduce propositional contents.[2] Illocutionary connections (YA-nodes) can be either anchored (associated, assigned) in locutions or in transitions. In the first case (see e.g. asserting, challenging, questioning), the locution provides enough information to reconstruct illocutionary force and content. Illocutionary connections are anchored in a transition when we need to know what a locution is a response to and to understand an illocution or its content. AIFdb Corpora allows for operation with either an individual NodeSet, or any grouping of NodeSets captured in a corpus. By integrating closely with the OVA+ (Online Visualisation of Argument) analysis tool (Janier et al., 2014), AIFdb Corpora allows for the rapid creation of large corpora compliant with AIF.

AIFdb provides the largest publicly available dataset comprising multiple corpora of analysed argumentation; and in addition AIF works as an interlingua facilitating translation from other representation languages with both the IAC and Microtext corpora in AIF format, for example. For both of these reasons, we have used AIF for our examples here (although the CASS technique itself is largely independent of annotation scheme).

[1] `http://www.aifdb.org`

[2] Illocutionary schemes are based on illocutionary forces defined in (Searle, 1969; Searle and Vanderveken, 1985).

S_1	20	18	29		39		31	18
S_2	20	17	17	12	27	12	31	18

Figure 1: Segmentation boundaries and mass for first and second annotators.

3.2 Comparing Analysis

Calculating the inter-annotator agreement of manual analysis, can be problematic when using traditional methods such as Cohen's kappa. In (Kirschner et al., 2015, p.3), the authors highlight this challenge: "as soon as the annotation of one entity depends on the annotation of another entity, or some entities have a higher overall probability for a specific annotation than others, the measures may yield misleadingly high or low values. (...) Therefore, many researchers still report raw percentage agreement without chance correction."

In the comparative statistics module we look to extend the solution in (Kirschner et al., 2015) in seven ways, by: (i) Calculating the segmentation differences between two annotations; (ii) Calculating propositional content relations using confusion matrices, accounting for all the nodes within an argument map and accounting for a differing segmentation; (iii) Calculating dialogical content relations (if they are contained in an argument map) using confusion matrices, accounting for all the nodes within an argument map and accounting for a differing segmentation; (iv) Defining the CASS technique to allow calculation scores to be combined; (v) Allowing the use of any metric for the CASS technique, which uses a confusion matrix, to give consistency to the area of argument mining; (vi) Providing results for not just inter-annotator agreement, but also, the comparison of manually annotated corpora against corpora automatically created by argument mining; (vii) Allowing the comparison of analysis given in different annotation schemes but migrated to AIF (e.g. compare text annotated in IAC to the annotation scheme from the Microtext corpus).

4 Comparative Statistics: Segmentation

Comparative statistics can provide for a number of cases with two main motivations: evaluation of automatic annotations against manual gold standards, and comparison of multiple manual annotations. The calculation is given between two separate annotations[3] A_1 and A_2 available in two separate corpora in AIFdb.

To account for a differing segmentation which does not doubly penalise the argument structure, the agreement calculation involves smaller subcalculations which can give an overview of the full agreement between annotators. Segmentation agreement considers the number of possible segments on which two annotators agree. A segmentation which differs between annotations can have a substantial effect on argument structure, such as the assignment of relations between proposition. An example is provided in Figure 1where segmentation is given for a first annotator (S_1) and a second annotator (S_2). In this case the two annotations give segments which resemble very similar mass(the number of words in a segment), however, more boundaries are placed in S_2 when compared to S_1 with a difference in granularity and a boundary misplaced by a word.

Three techniques are provided to tackle this problem with each recognising that a near miss (two segments that differ by a small margin, e.g. a word) should not be as heavily penalised as a full miss on the placement of segment boundaries. Performing the same calculation with $F1$ score or Cohen's kappa would result in a heavily penalised segmentation.

The P_k statistic (Beeferman et al., 1999), involves sliding a window of length k (where k is half the average segment size) over the segmented text. For each position of the window, the words at each end of the window are taken and the segment in which they lie is considered.

The WindowDiff statistic (Pevzner and Hearst, 2002), takes into account situations in which P_k fails. In P_k false negatives are penalised more than false positives, thus the agreement value could be unfair. The WindowDiff statistic remedies this by taking into account the number of reference boundaries and comparing this to the number of hypothesised boundaries.

The segmentation similarity statistic (S)(Fournier and Inkpen, 2012), again takes into account perceived failings of the P_k and WindowDiff statistics. Where both WindowDiff and P_k use fixed sized windows, which can adversely

[3]Throughout this paper A is used to denote annotation, l denotes a locution, p a propositional content node, ta a transition node and ra a propositional content relation.

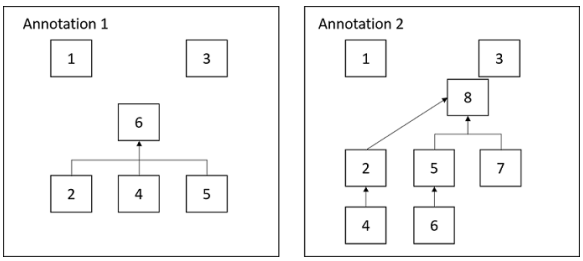

Figure 2: Propositional Content Relations for an annotation from AIFdb with first and second annotators.[4]

affect the outcome of an agreement calculation, *S* proposes that a minimum edit distance, scaled to the overall segmentation size, is considered. This edit distance allows near misses to be penalised but not to the same degree as a full miss.

5 Comparative Statistics: Propositional Relations

To compare relations it is important to calculate the agreement between each of the individual items which are annotated within an argument analysis. By providing calculations for individual items in an annotation we take into account that segmentation's may differ but do not penalise on this basis.

In the case of analysis with a differing segmentation, we use a guaranteed matching formula. This formula makes use of the Levenshtein distance (Levenshtein, 1966), where each locution or proposition in an annotation is compared with every locution or proposition in a second annotation. The Levenshtein distance for each comparison is taken and normalised, this is extended by using the position of words within the annotations taken from the original text. The position of words in the original text is important to correctly match propositions and locutions and therefore a proposition or locution which does not have matching positional words cannot be a match. In this situation the Levenshtein distance is increased (moved to zero) to account for a non-match. Each calculation taken is then stored in a matrix. The matrix is then traversed to find the smallest distance (highest value between zero and one), selecting the pair of locutions or propositions. This is continued until all nodes are matched or there are no matches which can be made, thus giving a Pareto optimal solution, a solution for which any match between propositions and locutions makes those individual matches consistent without making any other match worse and vice-versa.

An agreement calculation is given for all propositional content relations (support and attack relations). This calculation is based on the location of support and attack nodes within an analysis and the nodes to which they are connected. For a full agreement between annotators, a support or attack node must be connected between two propositions p_i, p_j, with these propositions being a match in A_1 and A_2. A support or attack node also has full agreement when one annotation is more fine grained but holds the same propositional content as the other annotation. For example, if annotation A_1 contains a support node which begins its relation in p_{bc} and gives a relation between p_{bc} and p_a, then this is the same as if A_2 had a support node with two separate propositions, p_b and p_c and related to p_a. This notion is extended when considering Figure 1 and Figure 2.

The differing segmentation in Figure 1 has an effect on the comparison between propositions. When considering propositions, non-identical propositions lead to a near zero similarity on support relations between these annotations. This is however an unintuitive approach to take, as the overall argumentative structure is penalised doubly (if we consider the segmentation and argumentative structure as different tasks) by the differing segmentation.

This is demonstrated in Figure 2 where the two annotators agree that there is a convergent argument been nodes four and five in annotation 1 and nodes five and seven in annotation 2. Extending this is proposition two of both annotations, where in annotation 1, proposition two is connected to four and five by a convergent argument. Yet in annotation 2, proposition 2 is a separate support relation. In the first instance of a convergent argument, splitting the segmentation calculation from the propositional relation calculation gives a fair representation of the argument structure without penalising for segmentation doubly.

In the second instance of a convergent argument and a separate support relation, there is a slight disagreement between the annotators. Despite both annotators agreeing that proposition two connects to the same node (propositions six in annotation 1 and eight in annotation two being the

[4]Numbered nodes represent propositions in the overall text and arrows represent support relations.

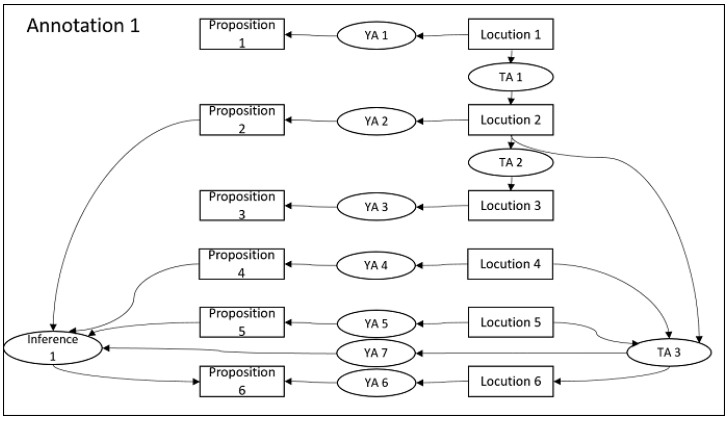

Figure 3: Full AIF IAT diagram for the annotation from the first annotator.

same node) a disagreement is shown because of the connection type if we consider Cohen's kappa or $F1$ score purely. Two options are available when calculating the similarity for this situation, either a similarity of zero is given or two separate calculations could be used with agreement on a premise conclusion basis but no agreement on the type of argument, thus giving a penalty.

To provide a confusion matrix all the possible node pairs to which a propositional content relation could be connected have to be considered. Any node pairs which both annotators have not connected are then counted and all nodes which are matched are counted, giving the observed agreement. All node pairs which the annotators do not agree upon are also counted.

6 Comparative Statistics: Dialogical Relations

Dialogical relations consider only the dialogue of an argument with the intentions of the speaker noted. A differing segmentation in various analysis can lead to low kappa or $F1$ scores. By splitting dialogical relations into a separate calculation it removes the double penalty assigned by segmentation. When comparing dialogical relations again we use the Levenshtein distance as described in Section 5.

A calculation is provided for illocutionary connections (YA) anchored in TAs or in Locutions. This calculation involves multiple categories, meaning a multiple category confusion matrix, due to the large number of possible YA-node types which can be chosen by annotators. An agreement is observed when both annotators select the same illocutionary connections. When A_1 contains a YA-node which is anchored in l_i and when A_2 contains the same YA anchored in l_i, then an agreement is observed. This also holds for TA's. The overall calculation then involves a confusion matrix where all disagreements are observed when YA nodes do not match. If we consider Figures 3 and 4 we can see between both annotations that there are a difference of four YA nodes.

A second calculation for YA-nodes, checking the agreement on the propositional content nodes in which they anchor and to where they are anchored (locution or TA), is also given. This calculation involves a multiple category confusion matrix. An example of when agreement is observed is when A_1 contains p_j anchored in a YA and the YA anchored in l_i and in A_2 the same structure with p_j and l_i is observed with the same YA node. The multi-category confusion matrix is calculated with disagreements observed when propositions and locutions do not match. When considering Figures 3 and 4 we see an example of agreement between the annotators on propositions 1, 2 and 3. Proposition 4, in Figure 3 and proposition 5, in Figure 4 also match and the same for proposition 5, in Figure 3 and proposition 7, in Figure 4. Disagreements are then observed with propositions 4 and 6 in Figure 4.

Three separate calculations are also given for TA-nodes. The first concerns the position of a TA node within locutions. Agreement is observed when A_1 contains a TA which is anchored in l_i and anchors l_j and A_2 contains the same TA anchored in l_i and anchoring l_j. For the final calculation all possible locution pairs are considered to give values for agreements on TA placement, agreements on non-TA placement and disagreements on TA placement. In the examples Figures 3 and 4 there

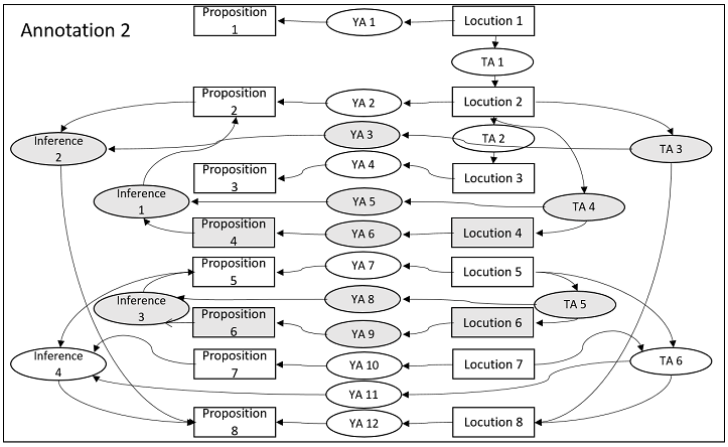

Figure 4: Full AIF IAT diagram for the annotation from the second annotator. Differences from Figure 3 are highlighted.

is a agreement between the annotators on TA 1, 2 and TA 3 in Figure 3 and TA 6, in Figure 4. A second calculation is then given for pairs of propositional content nodes and TA-nodes. When p_j is anchored in ta_i for A_1 and the same structure is observed in A_2 for the same propositional content node then there is agreement between the annotators. The overall confusion matrix is calculated by considering all pairs of TA-nodes and propositions and all disagreements between annotators. A third and final calculation is given for TA-nodes anchoring propositional content relations. For A_1 if ra_i is anchored finally in ta_i and ra_j is anchored finally in ra_j, in A_2 then agreement is observed. The overall confusion matrix is calculated by considering all possible pairs of TA's and propositional content relations. In Figures 3 and 4 agreement is observed only on inference 1 in Figure 3 and inference 4 in Figure 4. This provides a small penalty between the annotations for the added inference 2 in Figure 4, where earlier in Section 5 no penalty was given.

7 Aggregating into the CASS technique

Sections 4, 5 and 6 provide calculations for segmentation, propositional content relations and dialogical content relations. We have defined CASS which incorporates all of these calculation figures to provide a single figure for the agreement between annotators or a manual analysis and an automatic one, using both propositional content relations and dialogical content relations.

$$M = \frac{\sum P + \sum D}{n} \quad (1) \qquad CASS = 2\frac{M * S}{M + S} \quad (2)$$

In equation 1 the arithmetic mean, M, is the the sum of all propositional content calculations, P, plus the sum of all dialogical content calculations, D, over the total number of calculations made, n. We use this figure along with the segmentation similarity score to perform the harmonic mean and provide an overall agreement figure normalised and taking into account any penalties for segmentation errors. Equation 2 gives the CASS technique as the arithmetic mean, M, combined with the segmentation similarity, S.

The CASS technique allows for any consistent combination of scores to be used as either the propositional content calculations or dialogical calculations. That is to say that the CASS technique is not solely dependent on Cohen's kappa, or $F1$ score and can instead be substituted for any other overall measure. For the purpose of this example we will use the Cohen's kappa metric, as both annotations were annotated manually. We also use the S statistic for segmentation similarity as it handles the errors in P_k and WindowDiff statistics more effectively.

We sum both kappa scores giving an arithmetic mean, M, of 0.43. The S score, 0.95, is then combined with M in equation 2 to give an overall CASS of 0.59. scores this gives a fair representation of the overall agreement between the two annotators. In Table 1 the CASS technique is compared with Cohen's kappa and $F1$ score, where both scores do not take into account the slight difference in argument structure and therefore penalise this.

Method	Overall Score
Cohen's κ	0.44
CASS-κ	0.59
$F1$ score	0.66
CASS-$F1$	0.74

Table 1: Scores are provided for Cohen's kappa and $F1$ score, for both segmentation and structure, and CASS with S for segmentation and both kappa and $F1$ for structure.

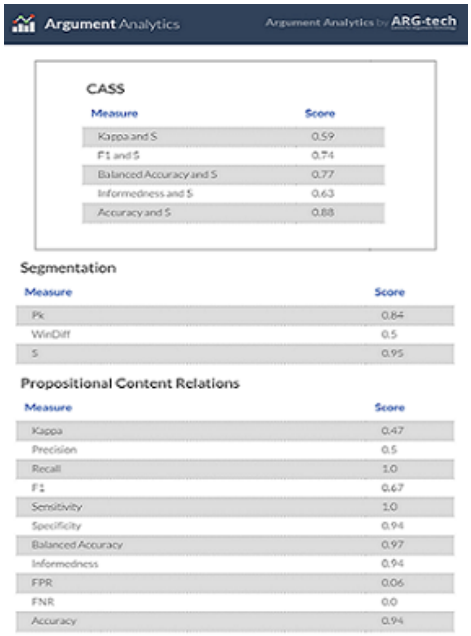

Figure 5: Screenshot of the comparative statistics module within Argument Analytics.

7.1 Extending Relation Comparisons

The CASS technique and comparative statistics module caters for the creation of confusion matrices for each calculation, allowing for the adaption of the overall results. This allows kappa, accuracy, precision, recall and $F1$ score all to be calculated, but, other metrics can also be considered for evaluating automatic analyses when using the CASS technique. Balanced accuracy (Brodersen et al., 2010), allows the evaluation of imbalanced datasets. When one class is much larger than the other Balanced Accuracy takes this into account and lowers the score appropriately. Informedness (Powers, 2011), gives the probability that an automatic system is making an informed decision when performing classification. A select set of metrics are part of the comparative statistics module, although, no metric is ruled out from this, allowing any metric employing a confusion matrix to use the CASS technique.

7.2 Deployment

Comparative statistics (see Figure 5) is part of the Argument Analytics suite which is to be publicly accessible at `http://analytics.arg.tech/`. It provides a suite of techniques for analysing sets of AIF data, with components ranging from the detailed statistics required for discourse analysis or argument mining, to graphic visual representations, offering insights in a way that is accessible to a general audience. Modules are available for: viewing simple statistical data, which provides both an overview of the argument structure and frequencies of patterns such as argumentation schemes; dialogical data highlighting the behaviour of participants of the dialogue; and real-time data allowing for the graphical representation of a developing over time argument structure.

8 Conclusions

Despite the widespread use of Cohen's kappa and $F1$ score in reporting agreement and performance, they present two key problems when applied to argument mining. First, they do not effectively handle errors of segmentation (or unitization); and second, they are not sensitive to the variety of structural facets of argumentation. These two problems lead to kappa and $F1$ underestimating performance or agreement of argument annotation.

The CASS technique allows for the integration of results for segmentation with those for structural annotation yielding coherent confusion matrices from which new CASS-κ and CASS-$F1$ scores can be derived. CASS is straightforward to implement, and we have shown that it can be included in web-based analytics for quickly calculating agreement or performance between online datasets. CASS offers an opportunity for increasing coherence within the community, aiding it to emulate the academic success of other subfields of computational linguistics such as summarization; and its subsequent deployment offers a simple way of applying it to future community efforts such as shared tasks and competitions.

Acknowledgments

We would like to acknowledge that the work reported in this paper has been supported in part by EPSRC in the UK under grants EP/M506497/1, EP/N014871/1 and EP/K037293/1.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, et al. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 84–93.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3121–3124.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 29–34.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161. Association for Computational Linguistics.

James B Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Nancy L Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 12–21.

Ivan Habernal and Iryna Gurevych. 2016. Argumentation mining in user-generated web discourse. *Computational Linguistics*.

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)*, pages 463–464, Pitlochry. IOS Press.

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 35–38.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 1–11.

Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)*, page 43.

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

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.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Marie-Francine Moens. 2013. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*.

Huy V Nguyen and Diane J Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 22–28.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the Second Workshop on Argumentation*

Mining. Association for Computational Linguistics, pages 116–126.

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 39–44.

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.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. *Proceedings of the First Conference on Argumentation, Lisbon, Portugal, June. to appear*.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

D.M.W. Powers. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2:27–63.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating toulmin model argumentation. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 45–55.

C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 56–66.

John R. Searle and Daniel Vanderveken. 1985. *Foundations of illocutionary logic*. Cambridge University Press.

John R. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 67–77.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. 2015. Learning sentence ordering for opinion generation of debate. In *Proceedings of the Second Workshop on Argumentation Mining. Association for Computational Linguistics*, pages 94–103.