Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines

Milagro Teruel[1], **Cristian Cardellino**[1], **Fernando Cardellino**[1], **Laura Alonso Alemany**[1], **Serena Villata**[2]
[1]Universidad Nacional de Córdoba, Argentina
[2]Université Côte d'Azur, CNRS, Inria, I3S, France
mteruel@unc.edu.ar, ccardellino@unc.edu.ar, fndcardellino@gmail.com, lauraalonsoalemany@unc.edu.ar, serena.villata@inria.fr

Abstract

In this abstract we present a methodology to improve Argument annotation guidelines by exploiting inter-annotator agreement measures. After a first stage of the annotation effort, we have detected problematic issues via an analysis of inter-annotator agreement. We have detected ill-defined concepts, which we have addressed by redefining high-level annotation goals. For other concepts, that are well-delimited but complex, the annotation protocol has been extended and detailed. Moreover, as can be expected, we show that distinctions where human annotators have less agreement are also those where automatic analyzers perform worse. Thus, the reproducibility of results of Argument Mining systems can be addressed by improving inter-annotator agreement in the training material. Following this methodology, we are enhancing a corpus annotated with argumentation, available at `https://github.com/PLN-FaMAF/ArgumentMiningECHR` together with guidelines and analyses of agreement. These analyses can be used to filter performance figures of automated systems, with lower penalties for cases where human annotators agree less.

Keywords: Argument Mining, Annotation Guidelines, Reproducibility

1. Introduction and Motivation

Argument Mining tackles a very complex phenomenon, involving several levels of human communication and cognition. Due to this complexity, data-driven approaches require a huge amount of data to properly characterize the phenomena and find patterns that can be exploited by an automatic analyzer. However, only small annotated corpora are available, and moreover they cannot be used in combination because they are based on different theoretical frameworks or cover different genres.

In this abstract we present work in progress in building a corpus annotated with arguments. As inherent part of this work, we are applying a methodology for early detection of ill-defined annotation concepts. We detect those by inspecting annotated texts for sources of disagreement between annotators, and redefining the annotation scheme so that these disagreements are minimized. This has been a successful approach to improve guidelines, as in (Hovy et al., 2006).

In preliminary explorations, we have found that agreement-driven modifications in the annotation scheme improve an automatic analyzer. Our final objective is to find an annotation scheme that is a tradeoff between theoretically based concepts, application needs, stability of human annotation and performance of automatic analyzers.

2. Annotated Corpus

2.1. ECHR Judgments

Four human annotators have annotated 7 judgments from the European Court of Human Rights (ECHR) in English, obtained from the Court website[1], totaling 28,000 words. Approximately half of the words were annotated as belonging to an argument component, as can be seen in Figure 1.

One of the judgments was annotated by all 4 annotators and discussed collectively as training. In this annotation, agreement between judges was never lower than $\kappa = .54$. Then, two pairs of judges annotated two judgments independently, we analyze agreement measures on those two pairs. More annotation pairs are currently being annotated and will be updated in the repository of the project (`https://github.com/PLN-FaMAF/ArgumentMiningECHR`).

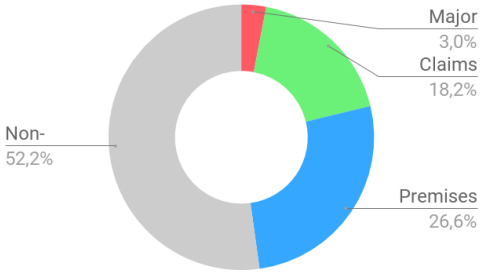

Figure 1: Proportion of component labels in the corpus.

2.2. Annotation Objectives

The objective of our annotation is to identify arguments composed by claims and premises (justifications) that are related to each other. Our initial annotation scheme was loosely based on (Toulmin, 2003), following the main adaptations that (Stab and Gurevych, 2015) propose to take the concepts from a theoretical model to practical annotation guidelines. Argument components were classified as:

Major Claim : a general statement expressing the author's stance with respect to the topic under discussion.

Claim : a controversial statement whose acceptance depends on premises that support or attack it. Claims are the central components of an argument and they either support or attack the major claim. We associate each claim with the actor that has issued it.

Premise : reasons given by the author for supporting or attacking the claims. They are not controversial but

[1]`hudoc.echr.coe.int`

factual. Specifically for this corpus, we distinguish the subclasses: Facts, Principles of Law and Case-law.

Argument components are connected to each other by relations, mainly **support** or **attack** relations (Simari and Rahwan, 2009). Claims support or attack other claims or a major claim, premises may support or attack claims or other premises. Additionally, we have established two more minor relations, specific for this corpus: **duplicate** (holding between claims or premises) and **citation** (holding between premises, when one cites a reference Case-law).
We have used brat (Stenetorp et al., 2012) as a tool for annotation.

3. Inter-annotator disagreement as an opportunity for improvement

3.1. Measuring disagreement

Disagreement between annotators is typically part of the annotation process in a qualitative way. Usually, annotation guidelines are iteratively refined in a long process where annotators discuss conflictive examples to specify vague concepts and establish annotation protocols.
Argument analysis is a highly subjective task, with typically low levels of inter-annotator agreement. Low inter-annotator agreement results in low reproducibility and also in poorer performance of automatic analyzers that are trained with these resources. We address reproducibility (and consequently automatic performance) by applying standard inter-annotator agreement measures to the annotated corpus from a very early stage, to make high-level decisions on the annotation scheme, instead of minute protocols (e.g., (Habernal, 2014)) to try to delimit concepts that have very high subject variation to begin with. We take low inter-annotator agreement as a symptom of ill-defined or far-fetched concepts, beyond the scope of our current annotation effort.
In order to assess the reproducibility of human annotations, we used Cohen's kappa (Cohen, 1960). This coefficient is a standard to measure inter-annotator agreement. It reports agreement between pairs of annotators, factoring out the probability that annotators would have agreed by chance. Other measures of inter-annotator agreement, like Krippendorf's alpha (Krippendorff, 1980) or Fleiss kappa (Fleiss and Cohen, 1973) will be included when we have a more extensive annotated corpus.

3.2. Annotators agree on what is argumentative

First of all, we found **high** agreement between annotators to determine whether a sentence contained an argument component, with Cohen's kappa ranging between $\kappa = .77$ and $\kappa = .84$. When this agreement is considered at token level, it varies between $\kappa = .59$ and $\kappa = .84$. We note that most disagreements occur between annotators that annotate less or more proportion of words as argumentative. Indeed, some annotators tend to consider more spans of text as argument components than others. However, there is a high agreement on spans identified as argumentative by annotators that consider less spans of text as argumentative. This has been addressed in the second version of the guidelines with a more detailed definition of argumentative text.

3.3. Major claims are major disagreements

For the classification of argument components as premises, claims or major claims, we found lower agreement, ranging from $\kappa = .48$ to $\kappa = .56$. Looking at the confusion matrices of annotations of pairs of annotators, in Figure 2, we find that there are important disagreements between all of the categories. However, the category of *major claim* seems to be the most conflictive: in one of the pairs, annotators did not have any overlap, in the other, they had more proportion of disagreement than of agreement. Therefore, this category seems to be ill-defined. Spans that are classified as major claims by one annotator tend to be classified as claims by the other, so we decided to collapse those two categories. When we do that, we obtain better agreement, as can be seen in Figure 3. We could think that this improvement is due to a smaller number of categories. However, the kappa coefficient, which factors out the number of categories, also improves: when those two categories are collapsed, then the agreement increases from $\kappa = .48$ to $\kappa = .51$ and from $\kappa = .56$ to $\kappa = .64$.

3.4. What the Court says are premises or claims

To analyze disagreements between premises and claims, we carried out a detailed analysis by subclasses, displayed in Figure 4. We found that claims issued by the ECHR are a major source of disagreement, because the concept is mixed with that of fact or principle of law. This can be expected, as claims by a Court in a judgment are performative speech acts that have the status of principles of law after the judgment is issued, and principles of law have the same status as facts in a reasoning by a court. However, epistemologically these three concepts are difficult to reconcile. To a minor extent, claims issued by the government tend to be mixed with premises labeled as facts, probably also because the legal status of the government, which can be easily assimilated to an actor capable of doing performative speech acts. Moreover, the category of premise as fact also accumulates a high number of disagreements with the category of non-argumentative text.
However, not considering fact premises as part of the annotation comes unnatural and is inadequate from a point of view of argumentation theory and of application utility. Therefore, in the second version of the guidelines, we have addressed this problem by refining the protocol, describing this phenomenon at length, and determining that the claims issued by the Court are to be taken as claims in the judgment where they are issued and as Principles of Law or Case Law if they are cited from another case.
In general, there is some confusion between premises interpreted as facts or as case-law, and also between premises considered case-law or law principles. However, these confusions can be easily addressed by a formal delimitation of case-law using shallow textual cues, also refining annotation guidelines.

3.5. Annotators seem to agree on relations

To assess the level of agreement for relations, we looked into relations that held between argument components where two annotators agreed. That meant between 46% and 74% of the components. For those, annotators agreed

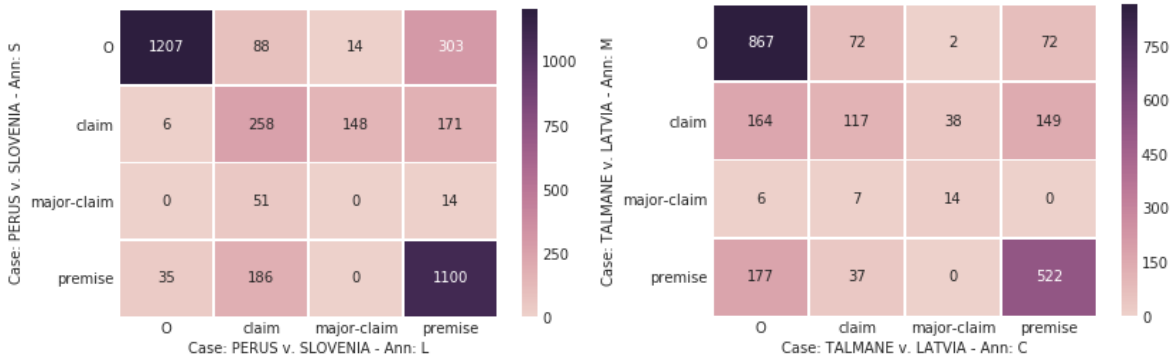

Figure 2: Confusion matrices for annotations of components between pairs of annotators, distinguishing major claims, claims and premises. Agreement for the matrix on the left is $\kappa = .56$ and on the right $\kappa = .48$.

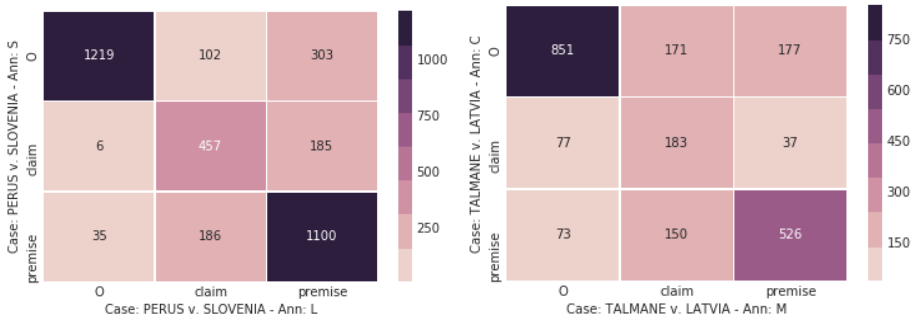

Figure 3: Confusion matrices for annotations of components between pairs of annotators, distinguishing only claims and premises. Agreement for the matrix on the left is $\kappa = .64$ and on the right $\kappa = .51$.

on the existence of a relation between components only in between 10% and 19% of the cases. When they agreed that a relation held between a given pair of components, annotators tended to agree on whether the relation was of attack, support or citation, with agreement ranging from 85% to 100% in most cases. However, the number of cases where such analysis could be carried out is so small that we require a bigger corpus to obtain more significant figures and draw conclusions upon them.

4. Automatic classification fails where humans disagree

In this section we show the relation between inter-annotator agreement and the performance of an automated classifier. To do that, we rely on the Argument classifier developed by (Eger et al., 2017), a neural end-to-end argumentation mining system with a multi-task learning setup. This system has been trained with part of the corpus, then annotated a different part of the corpus and its predictions compared with human annotations.

The comparison of human and automatic annotations is shown in Figure 5, with results showing the predictions of the classifier trained with major claims and not trained with major claims. We find that indeed major claims cannot be recognized by the classifier. This can be explained by the low proportion of major claims in the annotated corpus (see Figure 1), but neural classifiers tend to overfit the data and it could be expected that some major claims would have been identified. We also see that the confusion between premises and non-argumentative text is higher than the confusion between claims and non-argumentative text, and the confusion between premises and non-argumentative text is also higher than the confusion between claims and non-argumentative text. In consequence, there seems to be a strong relation between disagreements between humans and misperformance of automatic analyzers. Addressing the first will probably have a positive impact on the second.

5. Conclusions and Future Directions

We have presented an annotation effort for argumentation exploiting inter-annotator agreement as an indicator of ill-delimited concepts. Our aim is to enhance the reproducibility of argument annotation. Finding sources of disagreement across categories has allowed to make high-level decisions concerning the objectives of annotation. We have dropped the category of *major claim*, a major cause of disagreement that was not central to descriptive adequacy or application needs. We have described with more insight the case when the Court issues a claim that can be later considered as law, providing more insight and an unambiguous protocol of when these statements must be considered claims (when they are in the same judgments) or when they must be considered premises (when they were issued in previous judgments).

We have carried out preliminary experiments with a classifier, showing that automatic analyzers tend to fail where human annotators disagree.

In future work we will analyze the performance of the

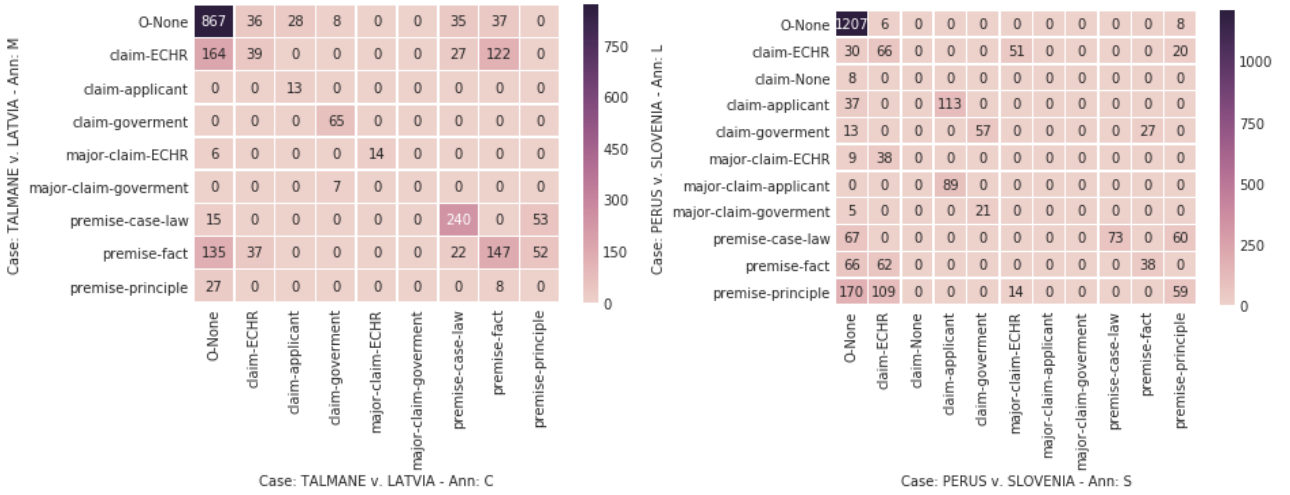

Figure 4: Confusion matrices for annotations of components between pairs of annotators, distinguishing their attributes. Agreement for the matrix on the left is $\kappa = .45$ and on the right $\kappa = .33$.

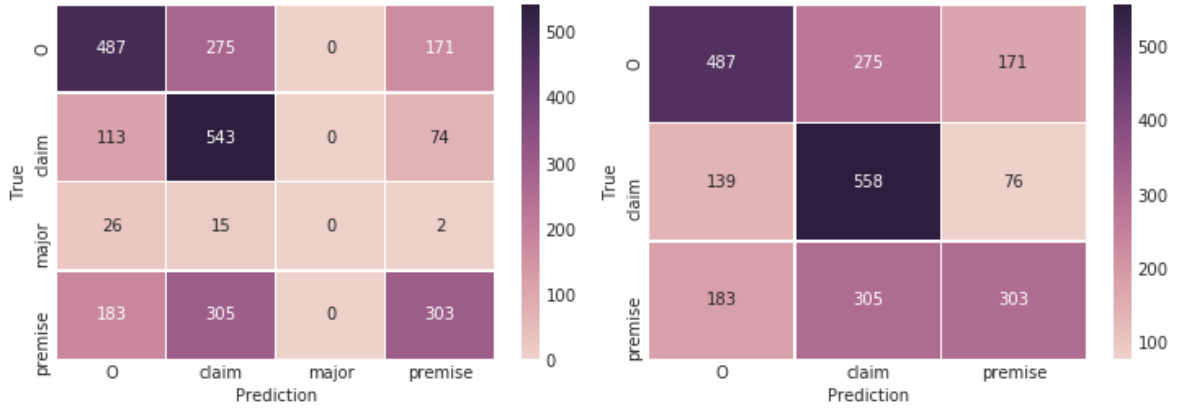

Figure 5: Confusion matrices for annotations of components between an automatic classifier and the human gold standard, distinguishing major claims (left) and not distinguishing them (right).

classifier with different configurations of the annotated resource: removing cases with low inter-annotator agreement and collapsing categories with high confusion between annotators. We will also be exploring inter-annotator agreement in relations.

We are currently annotating more documents with the second version of the guidelines. The newly annotated documents and the updated guidelines will be made available at the resource repository `https://github.com/PLN-FaMAF/ArgumentMiningECHR`.

6. References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psycological Measure*, 20:37–46.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. *CoRR*, abs/1704.06104.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Habernal, I., (2014). *Argumentation in User-Generated Content: Annotation Guidelines*. UKP Lab, Computer Science Department, Technische Universität Darmstadt, April.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90 In *NAACL 2006*, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krippendorff, K. (1980). *Content analysis: an introduction*. Sage, Beverly Hills, California.

Guillermo Ricardo Simari et al., editors. (2009). *Argumentation in Artificial Intelligence*. Springer.

Stab, C. and Gurevych, I., (2015). *Guidelines for Annotating Argumentation Structures in Persuasive Essays*. UKP Lab, Computer Science Department, Technische Universität Darmstadt.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *EACL 2012*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press, July.