Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 1–9
Marseille, 20 June 2022

Disagreement space in argument analysis

Annette Hautli-Janisz[1], **Ella Schad**[2], **Chris Reed**[2]
[1]Department of Computer Science and Mathematics, University of Passau
[2]Centre for Argument Technology, University of Dundee, UK
firstname.lastname@uni-passau.de
{ella, chris}@arg.tech

Abstract

For a highly subjective task such as recognising speaker intention and argumentation, the traditional way of generating gold standards is to aggregate a number of labels into a single one. However, this seriously neglects the underlying richness that characterises discourse and argumentation and is also, in some cases, straightforwardly impossible. In this paper, we present QT30nonaggr, the first corpus of non-aggregated argument annotation. QT30nonaggr encompasses 10% of QT30, the largest corpus of dialogical argumentation and analysed broadcast political debate currently available with 30 episodes of BBC's 'Question Time' from 2020 and 2021. Based on a systematic and detailed investigation of annotation judgements across all steps of the annotation process, we structure the disagreement space with a taxonomy of the types of label disagreements in argument annotation, identifying the categories of *annotation errors, fuzziness* and *ambiguity*.

Keywords: broadcast political debate, argumentation and conflict, Question Time, Inference Anchoring Theory

1. Introduction

State-of-the-art research in Natural Language Processing, in particular in areas like discourse parsing and argument mining, crucially relies on manually labelled data in order to be able to derive well-motivated computational models. However, labeling arguments and speaker intentions in natural language dialogue are tasks that are highly subjective: judgements are based on the knowledge of the topic under discussion, the speakers involved in the debate and their background. But even more so, it is the language of argumentation and debate that sets the challenge, independently of the underlying theory of argumentation, the annotation granularity and experience levels of the annotators.

Stab and Gurevych (2014) are the first ones to explicitly state that it is "hard or even impossible to identify one correct interpretation" of a particular argument structure, an issue confirmed by Lauscher et al. (2018) and Lindahl et al. (2019). Example 1 illustrates the issue based on an excerpt from our own data: Chika Russell, in a BBC's 'Question Time' on 8 July 2021, makes a comment in the context of UK local elections in 2021. The underlined part is argumentative and has been analysed with significantly different argument structure, provided in Figure 1: On the left-hand side, we find a serial structure including a rephrase ('Default Rephrase') and an inference ('Default Inference'), the right-hand side uses different segmentation and only has the propositional relation of 'Default Rephrase'.

(1) Chika Russel: *I have a view on how the election has gone. Call me Mystic Meg, if you will, people feel really forgotten, they feel let down. They feel the opportunities are not what they were.*

Although there is a clear understanding in the community that the analysis of argumentation is challenging due to a variety of factors, corpus development in this area is done in the "traditional" way: judgements are first collected based on guidelines for annotation, disagreements between labels are then resolved based on a (variety of) heuristics and eventually gold labels are assigned to argumentative units and the relations between them. These resolved labels then serve as the training base for a variety of machine learning techniques, with the significant drawback that they do not pass on the richness of information that is in fact encapsulated in the language.

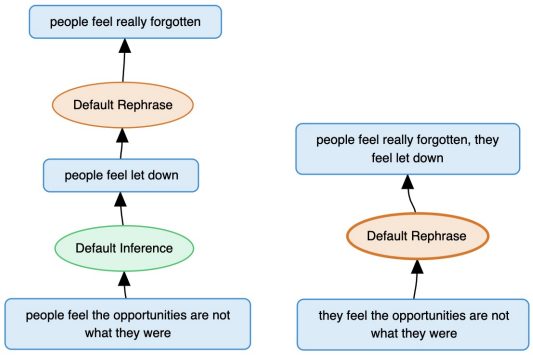

Figure 1: Two analyses for Example (1)

In this paper, we present QT30nonaggr, the first corpus of non-aggregated argument annotation. QT30nonaggr encompasses 10% of QT30, the largest corpus of dialogical argumentation and analysed broadcast political debate to date with 30 episodes of BBC's 'Question Time' from 2020 and 2021 (Hautli Janisz et al., 2022). Based on a systematic and detailed investigation of annotation judgements across all steps of the annotation process, we structure the disagreement space

with a taxonomy of the types of label disagreements in argument annotation, identifying the categories of *annotation errors, fuzzy language* and *ambiguity*. We therefore contribute a resource to the current and more general discussion of how computational models can be evaluated if more than one annotation label is available for an item, set out by (Uma et al., 2021). QT30nonaggr will also be the first non-aggregated corpus of argumentative data included in the Perspectivist Data Manifesto (`https://pdai.info/`).

2. Background

The core step in creating gold standards for supervised discourse parsing and argument mining is the resolution of multiple labels into a single "gold" label. This is done across all steps of manual argument analysis as defined by Lawrence and Reed (2020): text segmentation, argument/non-argument classification, simple argument structure and refined argument structure. The overwhelming number of approaches use the majority vote for deciding on a specific label (Rosenthal and McKeown, 2012; Stab and Gurevych, 2014; Wachsmuth et al., 2014; Hidey et al., 2017; Egawa et al., 2020). Habernal and Gurevych (2017) use majority voting, but employ adjudication in cases where majority is not possible (disagreement in segmentation leading to a different argument relation identification). Walker et al. (2012) use the mean rating across annotators for their labeling decision, Mochales and Moens (2011), Peldszus and Stede (2015) and Alliheedi et al. (2019) employ adjudication with an expert annotator to resolve label disagreements. Bar-Haim et al. (2020) use a threshold of 60% to judge whether an individual label is reliably annotated, judgements below that level are inconclusive which is claimed to be due to ambiguity. Toledo et al. (2019) and Gretz et al. (2020) take a number of annotator performance measures to discard what are presumed to be low-quality judgments. For cleansing argument data, (Dorsch and Wachsmuth, 2020) assume that in the case of indecisive annotations, the instance is kept in the dataset.

A more thorough investigation of the disagreement space for argument labeling is done to a significantly lesser extent: Stab and Gurevych (2014) investigate the disagreements encountered by way of confusion probability matrices for argument components and argumentative relations. They show that the major disagreement is between claims and premises and support/attack relations. Hidey et al. (2017) use an agreement matrix to show that disagreements are mostly between semantic types of claims, but they also note that ambiguity can lead to disagreements in segmentation and consequently argument structure. Habernal and Gurevych (2017) find that implicitness or topic relevance are relevant factors, Torsi and Morante (2018) show that segmentation, topic relatedness and commitment are crucial and Egawa et al. (2020) conclude that the majority of disagreement stems from semantic similarity.

The work presented in this paper deviates significantly from previous work: First, our annotation is not restricted to finding potentially isolated, but topic-relevant claims and relations. Instead, we label the complete debate with speaker intention and argumentation, including segments in which there is no argumentation, allowing us to derive how the debate unfolds. Secondly, we characterize the disagreement space along three dimensions which are implicitly (and sometimes explicitly) stated in related work: judgements go against the annotation guidelines (*Annotation Errors*), structures can be semantically and pragmatically fuzzy, i.e., judgements vary because the language is underspecified and leads to different interpretations (*Fuzziness*), and structure can be outright ambiguous, i.e., annotators pick up clearly separate interpretations based on syntactic, (lexical) semantic or pragmatic ambiguity (*Ambiguity*). Thirdly, we provide a non-aggregated resource for argumentative debate, QT30nonaggr, which serves as the basis for a large-scale investigation of the disagreement space in speaker intention recognition and argument analysis.

3. Inference Anchoring Theory (IAT)

Budzynska et al. (2014) and Budzynska et al. (2016) provide a theoretical scaffolding to handle *dialogue and argument structures, and the relations between them*, named Inference Anchoring Theory (IAT). The framework has been applied to over 2.5 million words in fifteen languages (available freely online at `corpora.aifdb.org`) and postulates three types of relations: (i) relations between content (propositional content of locutions); (ii) illocutionary connections that link locutions with their content and (iii) relations between locutions in a dialogue, called transitions. Given the scope of this paper, we only focus on the latter: relations between propositions, i.e., argumentative relations that hold between propositional content of speaker utterances.

3.1. Propositions

Propositions are derived from locutions and have the following properties: They are grammatical instantiations of the content of the locution. They have to be interpretable without context, i.e., they are standalone propositions that need to be intelligible without knowledge of surrounding propositional content. As a consequence, propositions may have to be reconstructed, so for instance elliptical or anaphoric expressions contained in the locution are resolved in the proposition. An example of this is the proposition of the third locution in Figure 2: 'they feel let down' (right-hand side) is resolved to 'people feel let down' in the proposition (left-hand side). The guideline for the annotators is to do minimal reconstruction in creating the proposition.

3.2. Propositional relations

Argumentative structures are relations between propositions; core IAT assumes three different relations that

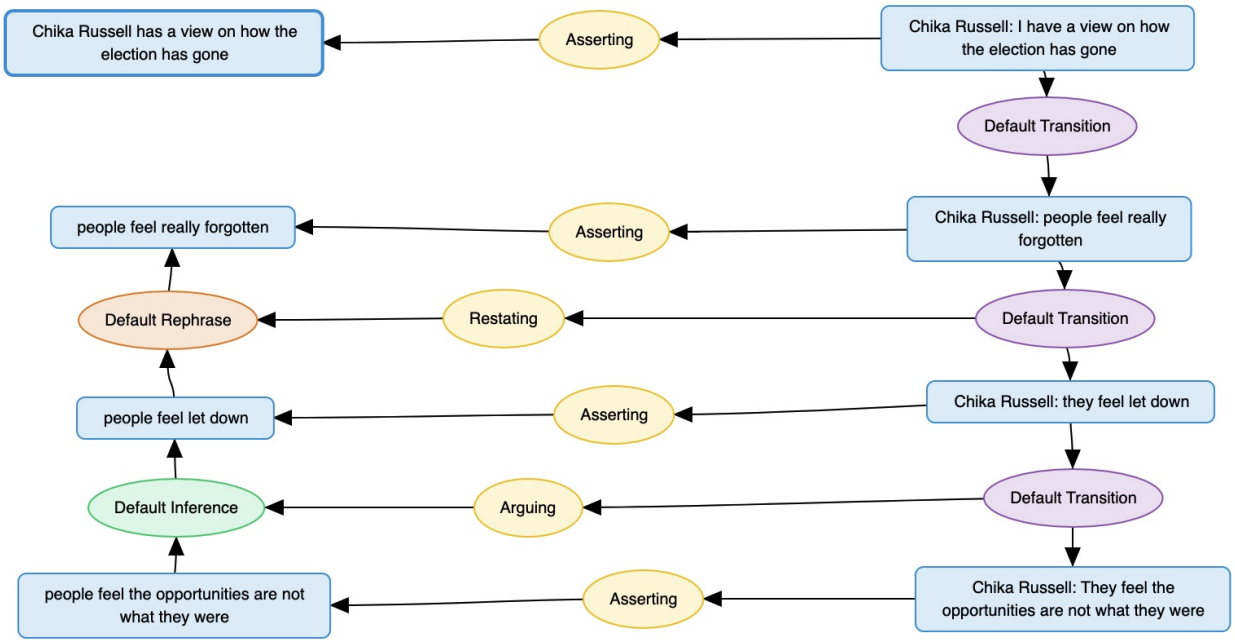

Figure 2: IAT diagram of Example (1), featuring locations (blue nodes on the right-hand side), propositions (blue nodes on the left-hand side), illocutionary relations (yellow nodes in the middle), dialogical relations (purple nodes on the right) and propositional relations – 'Default Inference' (green), 'Default Rephrase' (orange) and 'Default Conflict' (red).

are designed to capture argumentative structure in dialogue:

Inference (Support, 'Default Inference', RA, green node) Holds between propositions when one (or more) proposition is used to provide a reason to accept another proposition.

Conflict (Attack, 'Default Conflict', CA, red node) Holds between two propositions when one proposition is used to provide an incompatible alternative to another proposition.

Rephrase (Rephrase, 'Default Rephrase', MA, green node) Holds between two propositions when one proposition is used to rephrase, restate or reformulate another proposition. Rephrases also hold between questions and answers.

These relations are 'Default' in the sense that they can be instantiated with more specific relation types, for instance with presumptive argument scheme types (Walton et al., 2008).

Figure 2 provides the full IAT structure for one of the two analyses of Example (1), taken from the QT episode on 22 July 2021. Given the multiple layers of analysis needed for modeling dialogical argumentation, IAT graphs are divided into three different parts:

- The 'right-hand side' of nodes in the graph in Figure 2 is equivalent to a series of argumentative discourse units (Peldszus and Stede, 2013), i.e., the minimal unit into which the transcribed text is segmented. Each ADU has discrete argumentative function and records the name of the speaker in the form of 'firstname lastname : *locution content*'.
- The left-hand side encodes the propositional content of locutions, and the relationships of inference, conflict and rephrase between those propositional contents (as presented by the interlocutors).
- Those propositions are anchored in the dialogue via illocutionary connections (the 'middle', relations between locutions and propositions). These illocutionary acts capture speaker intention and are drawn from speech act theory (Searle, 1969; Searle and Vanderveken, 1985). There is a set of 10 relations that are used in IAT, namely *Asserting, Agreeing, Arguing, Assertive Questioning, Challenging, Disagreeing, Pure Questioning, Restating, Rhetorical Questioning* and *Default Illocuting*. For the scope of the current paper, however, we only focus on the left-hand side of the diagram: propositions and the argumentative relations between them.

3.3. Argument analysis

IAT analyses are produced with OVA+ (Online Visualisation of Argument – `http://ova.arg.tech/`), an open-source online interface for the analysis of argumentation in dialogues (Janier et al., 2014). OVA+ allows for a representation of the argumentative structure of a text as a directed graph. However, OVA+ does not

allow for an encoding of ambiguity or fuzziness, i.e., an annotator cannot indicate that she has identified a structure that licenses more than one solution. The guideline here is to pick the structure that is most likely given the speaker and the context. For compiling QT30nonaggr, we therefore use IAT graphs that are created by different annotators and compare their analyses.

These graphs are created in several steps which mostly correspond to those based on the steps for manual argument analysis set up by (Lawrence and Reed, 2020) and briefly described here:

Chunking The starting point for analysis is a text of around 10,000 words (+/- 20%) that is first chunked into (40-80) excerpts each comprising around 150-250 words – passages that are small enough to be considered in total by an analyst, but large enough to include substantial dialogical exchange. We exploit natural topical, thematic and turn-based breaks to guide this chunking process. This chunk is then passed to different annotators for analysis.

Segmentation In the first step, an analyst segments the text into (argumentative) discourse units (or 'locutions' in IAT), producing between five and thirty locutions per excerpt. Going back to Example 1, the analyst decides whether to split the excerpt in two or three locutions. In the same step, the analyst reconstructs information to be recorded in the proposition, for instance anaphora and ellipses (see §3.1).

Classification of units The analyst then identifies whether a locution has indeed argumentative function or not. A crucial decision factor here is speaker intention: was the speaker intending to make an argument here, also given the larger societal or political context. This decision can go hand in hand with the previous step of segmentation: an analyst might have different solutions to capturing the argumentative structure and makes a first decision by segmenting the text in a particular way, shown in Figure 1.

Structure identification After classification an analyst immediately adds the type of propositional relation and the illocutionary connection between locutions (right-hand side) and propositions (left-hand side). The result is a map containing between a dozen to a hundred nodes in total, depending on the length and the content of the excerpt.

Review Each analysis map then undergoes peer review by which a randomly chosen second analyst who reviews and discusses annotation choices with the first.

4. Data

As the basis of our investigation we use QT30, the largest corpus of analysed dialogical argumentation ever created (19,842 utterances, 280,000 words) and also the largest corpus of analysed broadcast political debate to date, using 30 episodes of BBC's 'Question Time' from 2020 and 2021 (Hautli Janisz et al., 2022). Question Time is the prime institution in UK broadcast political debate and features questions from the public on current political issues, which are responded to by a weekly panel of five figures of UK politics and society. QT30 is highly argumentative and combines language of well-versed political rhetoric with direct, often combative, justification-seeking of the general public. In total, the corpus features 10,818 propositional relations, i.e., argumentative structures. Inference ('supports') and Rephrase have the highest frequency, 48% and 42.6%, respectively. Conflicts are significantly less frequent, making up only 9.4% of all relations between propositions. The resource is freely available at `http://corpora.aifdb.org/qt30nonaggr`.

The annotation was conducted by 38 students of linguistics, philosophy, literature and computer science in Scotland, England, Germany and Poland. More than 60 students took part in one of three rounds of training in 2020 and 2021. Topic of the 15 hour course (taught in person once in 2020 and then virtually three times in 2020 and 2021) was a general introduction to argumentation theory and detailed instructions on applying Inference Anchoring Theory to dialogical argumentation across genres. Due to the strict quality restrictions for QT30, only the top 38 annotators were selected to contribute.

The Combined Argument Similarity Score (CASS) (Duthie et al., 2016), which calculates separate scores for segmentation, argumentative structures and illocutionary forces and aggregates them into a single score for annotator agreement, for all of QT30 is 0.56, signaling moderate agreement. Despite the fact that other papers report slightly higher CASS scores – 0.752 in Visser et al. (2019) and κ = 0.75 in Budzynska et al. (2014)) – inter-annotator agreement for QT30 is based on a very heterogeneous but realistic dataset for quantifying annotation reliability: it features annotations by all 38 annotators which are based on a variety of experience levels due to the incremental formation of the annotation team.

Given the significant expertise level of the annotators, we hypothesize that the CASS score hints at more systematic annotation differences that go beyond simple annotation errors. Instead, we hypothesise that it hints at the deeper issue of subjectivity in discourse-level tasks such as argument analysis, manifested by the fuzziness and ambiguity of language and discourse in general. The different dimensions of labeling disagreements are elaborated on in the following.

5. A taxonomy of label disagreements

As the basis for our empirical investigation of label disagreements in argument analysis, we randomly select four excerpts of each episode (about 8-10% of QT30) and request a second annotation by a random other member of the annotation team. This second annotation is conducted in the standard procedure described in §3.3, review is done by another randomly assigned annotator. The annotators are not aware that they con-

tribute their analysis for the purpose of identifying labeling disagreements instead of regular corpus analysis.

For the empirical analysis of the disagreed-upon labels, one of the most senior analysts is manually investigating the two different graphs per excerpt in parallel. There were several loops in identifying an appropriate partitioning of the disagreement space, based on previous work and informed by the special patterns that dialogical argumentation is delivering. In the following we present the three dimensions that allow us to characterise the disagreement space for dialogical argument analysis, distinguishing the categories of *annotation errors*, structures of *fuzziness* and *ambiguity*.

5.1. Annotation errors

The first dimension of label disagreements are simple annotation errors that violate annotation criteria which are clearly stated in the annotation manual.[1] We illustrate those categories with clear-cut examples from the corpus.

Discourse-structuring material is retained (ERR-DISC) This category captures any discourse material that is wrongly retained in the proposition. A typical example is a discourse marker such as 'because', which is kept in the proposition despite the fact that its meaning is captured by the relation between propositions, e.g. 'Default Inference'.

Another example is the recording of epistemic markers like 'I think' in the proposition. Such material is again not a part of the asserted content proper, as it signals the knowledge or belief – the epistemic attitude – of the speaker towards the proposition at issue. (Occasionally the epistemically qualified proposition is precisely what is at stake – *No, no he'd didn't **think it** he **knew it*** – but this is an extreme exception.) A clear example (node #657575) is offered by (2) and its alternate (3) :

(2) I think the handling of the pandemic has shown there will be divides after it's over.

(3) the handling of the pandemic has shown there will be divides after it's over.

Grammar errors (ERR-GRAM) The annotation guidelines require every proposition to be grammatically well-formed. This particularly holds for syntactic structure, which is not always standard in natural occurring dialogue and needs to be corrected by the annotator. This may include extraneous punctuation, such as ellipses or periods, as well as wrongly capitalised letters or wrong conjugation.

Reconstruction errors (ERR-REC) Another key requirement of IAT analysis is that every proposition is intelligible without additional context. This means that for instance anaphora and elliptical constructions are resolved where possible. The guideline is to stay as close to the original text and only add minimal extra material to facilitate understanding. With ERR-REC we mark any proposition that is wrongly reconstructed, i.e., using a wrong referent. An interesting example is provided by the two reconstructions of (4) (node #838276) in (5) and (6). Here, the context does not license the reconstruction with the structure in (5), instead 'it' is resolved with full clause ('letting footballers and VIPS in without quarantine while lockdown is still in place').

(4) It is deep hypocrisy

(5) Robert Buckland is hypocritical

(6) letting footballers and VIPs in without quarantine while lockdown is still in place is deep hypocrisy

Erroneous relations (ERR-REL) This category captures annotation of propositional relations that violate the guidelines, for instance if the connection between a premise and a conclusion is pointing in the wrong direction, or – given that IAT is tracking the dialogical unfolding of argumentation – where a conflict is analysed from earlier to later discourse material (whereas only material that has already been introduced into the discourse can be the target of a conflict). Another common pattern of ERR-REL is erroneously marking question-answering – particularly because in political discussion, question-answering and question-avoiding is so common. An incorrect analysis of a response as an answer is shown in Fig. 3; the correct analysis is in Fig. 4, that highlights the fact that the response in fact provides no answer at all.

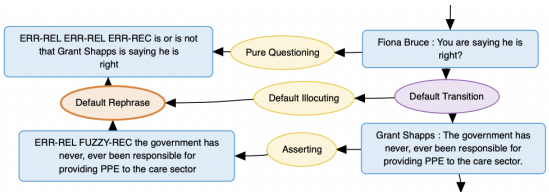

Figure 3: Incorrect annotation of question-answering in AIFdb map #23446

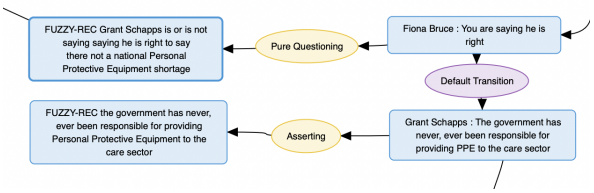

Figure 4: Correct annotation of non-question-answering in AIFdb map #23125

Erroneous splitting (ERR-SPLIT) The guideline for splitting text into segments clearly states that segments do not go beyond the sentence boundary. They

[1] http://www.arg.tech/f/IATannotationguidelines.pdf

also specify that any unit with discrete argumentative function has to be kept separately. For instance, the 'if...then' construction in Fig. 6 is an instance of a clear-cut inferential relation between the 'if'-clause and the 'then-clause' (despite them being inverted). The alternate analysis in Fig. 5 one misses the split and therefore the argumentative relation. In this case, we also mark ERR-REL on the wrongly-split proposition.

5.2. Fuzziness

This dimension of the disagreement space originates to some extent in the genre under investigation: natural, spontaneous argumentation features language patterns that are vague, fuzzy and therefore result in different analyses which themselves are valid, but illustrate the uncertainty in representing partially underspecified or vague language.

Fuzzy content (FUZZY-DISC) With this category we label all instances where the content of the locution is fuzzy in terms of whether parts of the locution serve as discourse-structuring material (which is not captured in the proposition) or contribute content that is (potentially) argumentatively relevant and therefore kept in the proposition. Fig. 7 shows a good example, in which the *winding me up* material has been analysed as a proposition (allowing it thereby to be referenced argumentatively later – "*yeah, it really winds me up too*" for example).

Fuzzy reconstruction (FUZZY-REC) This disagreement label is used in cases where the reconstruction of anaphoric or elliptical content varies across annotations. This is particularly the case for the reconstruction of 'that', where annotators judge the scope of the antecedent to be of varying length such as the different between the analyses in maps #22924 and #22930, (7) and (8), respectively. In some cases, annotators vary in their exact spell-out of the antecedent (though they mean the same entity), e.g., 'David Unknown' in node #843333 versus 'David Davies' in node #720703.

(7) leaving the European Union has or has not helped in speeding up the process of vaccine creation

(8) leaving the European Union has or hasn't helped in speeding up the vaccination delivery process

Fuzzy relation (FUZZY-REL) In this category we subsume all instances where the relation between the propositions ('Default Inference', 'Default Conflict' and 'Default Rephrase') is the same between two maps, however the splitting of argumentative units is slightly different. This can, for instance, mean that one analyst has chosen a linked argument structure (more than one premise leading to the conclusion, the premises are dependent on each other) versus a convergent argument (more than one premise to the conclusion, but the premises are independent of each other).

Fuzzy transcript input (FUZZY-TRANS) This category of disagreement is due to the data source under analysis: IAT analysis is conducted based on transcripts of natural dialogues and we do see cases in which the stenographer is not able to provide a clear recount of the conversation, for instance due to crosstalk or interruptions between interlocutors. This can lead to fragmented text which annotators treat differently in their analyses.

5.3. Ambiguity

The third dimension of disagreement captures ambiguous structures in the dialogue. In contrast to fuzzy language, we treat ambiguity as those instances where a string yields two fully discrete discourse or argumentative structures. In the following we briefly illustrate the different types of ambiguity that arise in the data:

Ambiguous anaphoric expressions (AMB-ANAPH) Given that annotators have to create propositions that are understandable without context, one core step of analysis is anaphora resolution. Similarly to (Poesio and Artstein, 2005), we also note the key challenge that the demonstrative 'that' poses for reconstruction. But it is also structures as in example 5.3, taken from a discussion on the Omicron wave in the episode on 1 July 2021:

(9) Andy Burnham: *I think I'm right in saying cases were highest day than they were in January. That's a worry. But you are right to say, Fiona, it isn't translating into hospitalisation. I was discussing the figures just before the show with David. So creeping up.*

The last sentence sentence contains an elliptical construction, which was resolved to 'deaths are creeping up' (map #23384) by one annotator, whereas the other analysis captures it as 'cases are creeping up' (map #23385). Both structures are discrete and correct and are therefore marked for ambiguity.

Ambiguous argument structure (AMB-REL) This category encompasses all analyses that exhibit two discrete argumentative structures. An example of this is given in Figure 1: Based on a different splitting decision, different argument structures arise: a serial argument with three propositions, connected by a 'Default Rephrase' and a 'Default Inference' (left-hand side) versus two propositions related by a single 'Default Rephrase'. Both analyses are valid given the context and are therefore labeled as ambiguous.

Ambigous splitting (AMB-SPLIT) Central to the analyses in Figure 1 and directly related to the previous category of AMB-REL is the category of ambiguous splitting, i.e., argumentative units have different length, but both segmentation decisions are well-motivated and adhere to the annotation guidelines.

In what follows, we briefly describe QT30nonaggr, the resource that is generated based on the analysis of the disagreement labels.

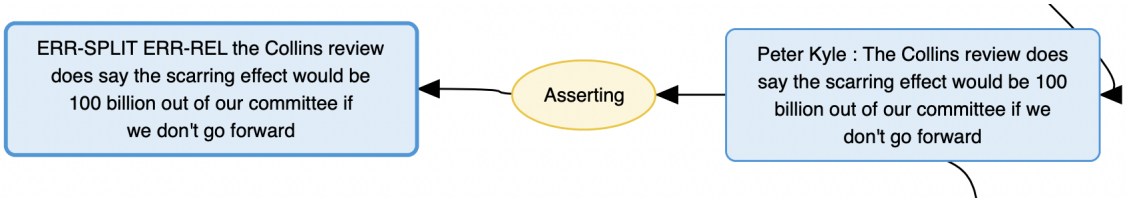

Figure 5: Incorrect splitting of if-then in AIFdb map #23298

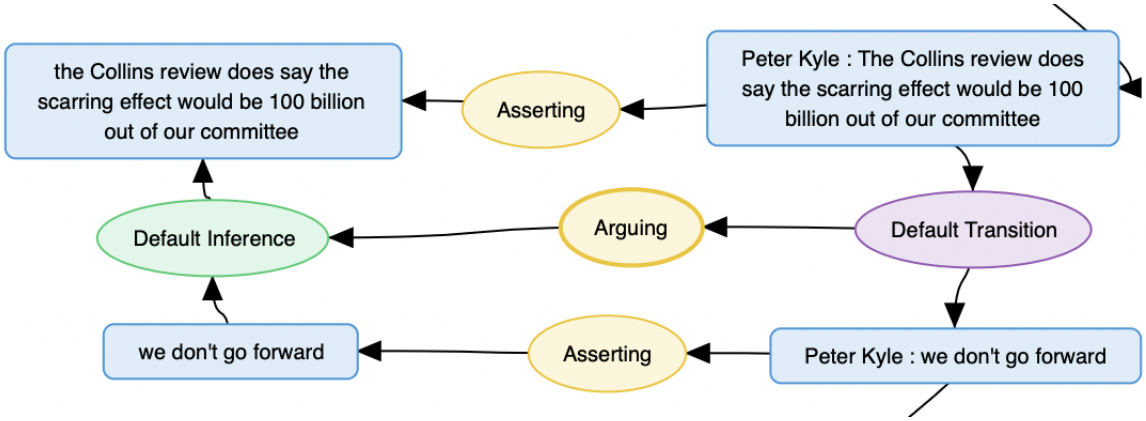

Figure 6: Correct splitting of if-then in AIFdb map #23290

6. QT30nonaggr

QT30nonaggr contains 67 excerpts which are annotated independently by two annotators (134 graphs in total). The length of excerpts ranges between 150-250 words. Overall, the resource contains 1817 propositions with an average length of 14.19 words per proposition. 'Default Inference' is the most frequent propositional relation (546), followed by 'Default Rephrase' (485) and 'Default Conflict' (106).

QT30nonaggr contains the full IAT graphs (as illustrated in Figure 2) plus the disagreement labels specified in §5. Identifying disagreements in illocutionary labels (the yellow connections in the middle of Figure 2), we leave for further work, however we tend to see significantly fewer disagreements there than in actual argument analysis.

Table 1 gives the detailed numbers for the disagreement space in QT30nonaggr: The category of annotation errors makes up the largest share of label disagreements by far – 907 out of 1402 (65%). Disagreements based on fuzzy language are second-most frequent (288/1402 – 20%), instances of ambiguity make up 207 out of 1402 disagreements (15%). Some disagreement labels appear in a vast majority of maps, the top ones being ERR-REL (92%), AMB-REL (85%) and FUZZY-REC and ERR-REC (both 82%). This confirms findings of previous work, e.g., (Stab and Gurevych, 2014), which shows that it is particularly the identification of relations that presents a challenge.

Label	% of graphs	# of labels
Errors		**907**
ERR-DISC	83%	130
ERR-GRAM	44%	46
ERR-REC	82%	258
ERR-REL	92%	355
ERR-SPLIT	71%	118
Fuzzy		**288**
FUZZY-DISC	41%	40
FUZZY-REC	82%	176
FUZZY-REL	49%	47
FUZZY-TRANS	27%	25
Ambiguity		**207**
AMB-ANAPH	21%	14
AMB-REL	85%	146
AMB-SPLIT	48%	47
Total		**1402**

Table 1: The detailed number for characterising the disagreement space of QT30nonaggr.

7. Summary

The analysis and reconstruction of argument is a challenging task. When taught as part of a critical thinking undergraduate programme, or in the context of study skills, or even in formal settings such as intelligence analysis or jurisprudence, it is well recognised that texts will support multiple interpretations. More recently, this has yielded particular challenges for the computational linguistics community, which naturally works from an assumed basis of a single, agreed-upon, gold standard. In our work constructing the largest corpora of annotated argument and debate currently available, we have encountered these challenges

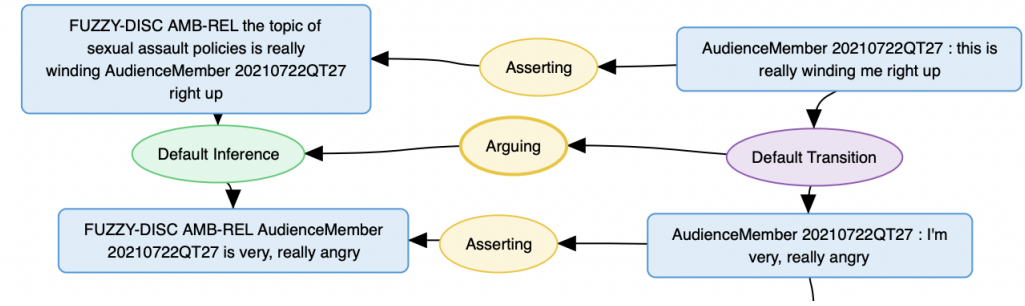

Figure 7: Material that is arguably discourse structuring in AIFdb map #23458

head-on, and have collated our experiences into a new non-aggregated corpus, QT30nonaggr, which not only documents cases of mismatching annotations, but also aims to provide an initial classification of the most prominent ways in which annotation discrepancies occur. Though as De Morgan famously said, "*There is no such thing as a classification of the ways in which men may arrive at an error: it is much to be doubted whether there ever can be,*" our approach here is to provide a starting point for exploring how errors might be arrived at both in annotating argumentation and reasoning structures, and, thereby in the long run, also in how errors are arrived at in general understanding of such structures.

8. Bibliography

Alliheedi, M., Mercer, R. E., and Cohen, R. (2019). Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, Florence, Italy, August. Association for Computational Linguistics.

Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., and Slonim, N. (2020). From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online, July. Association for Computational Linguistics.

Budzynska, K., Janier, M., Kang, J., Reed, C., Saint-Dizier, P., Stede, M., and Yaskorska, O. (2014). Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Budzynska, K., Janier, M., Reed, C., and Saint Dizier, P. (2016). Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Dorsch, J. and Wachsmuth, H. (2020). Semi-supervised cleansing of web argument corpora. In *Proceedings of the 7th Workshop on Argument Mining*, pages 19–29, Online, December. Association for Computational Linguistics.

Duthie, R., Lawrence, J., Budzynska, K., and Reed, C. (2016). The CASS technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 40–49, Berlin, Germany, August. Association for Computational Linguistics.

Egawa, R., Morio, G., and Fujita, K. (2020). Corpus for modeling user interactions in online persuasive discussions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France, May. European Language Resources Association.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. *AAAI 2020*, 34.

Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April.

Hautli Janisz, A., Kikteva, Z., Siskou, W., Gorska, K., Becker, R., and Reed, C. (2022). Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC2022)*. ACL.

Hidey, C., Musi, E., Hwang, A., Muresan, S., and McKeown, K. (2017). Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark, September. Association for Computational Linguistics.

Janier, M., Lawrence, J., and Reed, C. (2014). Ova+: An argument analysis interface. In *Computational Models of Argument: Proceedings of COMMA*, volume 266, pages 463–464.

Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018). An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium, November. Association for Computational Linguistics.

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

Lindahl, A., Borin, L., and Rouces, J. (2019). Towards assessing argumentation annotation - a first step. In

Proceedings of the 6th Workshop on Argument Mining, pages 177–186, Florence, Italy, August. Association for Computational Linguistics.

Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artifical Intelligence and Law*, 19:1–22.

Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, jan.

Peldszus, A. and Stede, M. (2015). An annotated corpus of argumentative microtexts. In D. Mohammed et al., editors, *Argumentation and Reasoned Action – Proc. of the 1st European Conference on Argumentation, Lisbon*. College Publications, London.

Poesio, M. and Artstein, R. (2005). Annotating (anaphoric) ambiguity.

Rosenthal, S. and McKeown, K. (2012). Detecting opinionated claims in online discussions. *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.

Searle, J. and Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. (2019). Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China, November. Association for Computational Linguistics.

Torsi, B. and Morante, R. (2018). Annotating claims in the vaccination debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56, Brussels, Belgium, November. Association for Computational Linguistics.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:385–1470.

Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2019). Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, Feb.

Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., and Palakarska, T. (2014). A Review Corpus for Argumentation Analysis. In Alexander Gelbukh, editor, *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014)*, pages 115–127, Berlin Heidelberg New York, April. Springer.

Walker, M., Tree, J. F., Anand, P., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.