Towards Feasible Guidelines for the Annotation of Argument Schemes

Elena Musi†, Debanjan Ghosh* and Smaranda Muresan†
†Center of Computational Learning Systems, Columbia University
*School of Communication and Information, Rutgers University
em3202@columbia.edu, debanjan.ghosh@rutgers.edu, smara@ccls.columbia.edu

Abstract

The annotation of argument schemes represents an important step for argumentation mining. General guidelines for the annotation of argument schemes, applicable to any topic, are still missing due to the lack of a suitable taxonomy in Argumentation Theory and the need for highly trained expert annotators. We present a set of guidelines for the annotation of argument schemes, taking as a framework the *Argumentum Model of Topics* (Rigotti and Morasso, 2010; Rigotti, 2009). We show that this approach can contribute to solving the theoretical problems, since it offers a hierarchical and finite taxonomy of argument schemes as well as systematic, linguistically-informed criteria to distinguish various types of argument schemes. We describe a pilot annotation study of 30 persuasive essays using multiple minimally trained non-expert annotators .Our findings from the confusion matrixes pinpoint problematic parts of the guidelines and the underlying annotation of claims and premises. We conduct a second annotation with refined guidelines and trained annotators on the 10 essays which received the lowest agreement initially. A significant improvement of the inter-annotator agreement shows that the annotation of argument schemes requires highly trained annotators and an accurate annotation of argumentative components (premises and claims).

1 Introduction

Argumentation is a type of discourse in which various participants make arguments, presenting some premises in support of certain conclusions, with the aim of negotiating different opinions and reaching consensus (Van Eemeren et al., 2013). The automatic identification and evaluation of arguments require three main stages: 1) the identification, segmentation and classification of argumentative discourse units (ADUs), 2) the identification and classification of the relations between ADUs (Peldszus and Stede, 2013a), and 3) the identification of argument schemes, namely the implicit and explicit *inferential relations* within and across ADUs (Macagno, 2014).

Although considerable steps have been taken towards the first two stages (Teufel and Moens, 2002; Stab and Gurevych, 2014; Cabrio and Villata, 2012; Ghosh et al., 2014; Aharoni et al., 2014; Rosenthal and McKeown, 2012; Biran and Rambow, 2011; Llewellyn et al., 2014), the third stage still constitutes a major challenge because large corpora systematically annotated with argument schemes are lacking. As noticed by Palau and Moens (2009), this is due to the proliferation in Argumentation Theory of different taxonomies of argument schemes based on weak distinctive criteria, which makes it difficult to develop intersubjective guidelines for annotation. In the *Araucaria* dataset (Reed and Rowe, 2004), for example, two argument scheme sets other than Walton's are used as annotation protocols (Katzav and Reed, 2004; Pollock, 1995).

To overcome this problem, the most successfully applied strategy has been to pre-select from existing larger typologies, such as that of Walton et al. (2008), a subset of argument schemes which is most frequent in a particular text genre, domain or context (Green, 2015; Feng and Hirst, 2011; Song et al., 2014; Schneider et al., 2013) and provide annotators with critical questions as a means to identify the appropriate scheme. Such a bottom up approach allows one to improve the identi-

Proceedings of the 3rd Workshop on Argument Mining, pages 82–93,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

fication conditions for a set of argument schemes (Walton, 2012), but it is hardly generalizable since it is restricted to specific argumentative contexts. Moreover, while critical questions constitute useful tools to evaluate the soundness of arguments (Song et al., 2014), they are far less suitable as a means to identify the presence of arguments: adopting a normative approach, annotators would conflate the notion of "making an argument" with that of "making a sound argument", while defeasibility should not be considered as an identification condition for the mere retrieval of arguments in texts.

We hypothesize that the *Argumentum Model of Topics* (Rigotti and Morasso, 2010; Rigotti, 2009), an enthymematic approach for the study of the inferential configuration of arguments, has the potential to enhance the recognition of argument schemes. Unlike other approaches (Van Eemeren and Grootendorst, 1992; Walton et al., 2008; Kienpointner, 1987), it offers a *taxonomic hierarchy* of argument schemes based on criteria which are distinctive and mutually exclusive and which appeal to *semantic properties of the state of affairs* expressed by premises/claims, and not to the logical forms (deductive, inductive, abductive) of arguments, whose boundaries are still debated (Section 2). However, even if these semantic properties are linguistically encoded, and hence potentially measurable, they might call for some background knowledge in frame semantics to be identified as well as for quite specific analytic skills. Moreover, the cognitive load requested by the annotation of argument schemes is higher than that needed for the annotation of the argumentative discourse structure (e.g., argument components such as claims and premises, and argument relations such as support/attack). As stated by Peldszus and Stede (2013b) with regard to the annotation of argument structure in short texts, the inter-annotator agreement among minimally trained annotators is bound to be low due to different personal commitments as well as interpretative skills of the texts. We wanted to test whether this conclusion is valid for our annotation task.

We conducted a pilot annotation study using 9 minimally trained non-expert annotators. As a corpus we used 30 short persuasive essays already annotated as to premises, claims and support/attack relations (Stab and Gurevych, 2014). Section 3 presents the set of guidelines and our study. Our findings from measuring the inter-annotator agreement (IAA) support previous findings that annotation of argument schemes would require highly trained annotators (Section 4). We also performed an analysis of confusion matrices to see which argument schemes were more difficult to identify, and which parts of the guidelines might need refinement (Section 4). Another finding of this study is that the identification of argument schemes constitutes a means to refine the annotation of premises and claims (Section 5). We refined the guidelines and tested them through the annotation of the 10 essays which received the lowest inter-annotator agreement using 2 trained non-expert annotators and 1 expert annotator (Section 6). The results show an improvement in the inter-annotator agreement. The confusion matrix suggests that the frequency of non-argumentative relations between premises/claims, claims/major claims highly affects disagreement. The guidelines and the annotated files are available at: `https://github.com/elenamusi/argscheme_aclworkshop2016`.

2 Theoretical Background and Framework

As Jacobs (2000, 264) puts it, "arguments are fundamentally linguistic entities that express [...] propositions where those propositions stand in particular inferential relations to one other". These *inferential relations*, namely argument schemes, are textually implicit and have to be reconstructed by the participants of a critical discussion in order to reach agreement or disagreement. In everyday life this happens quite intuitively on the basis of common ground knowledge: everyone would agree that "The sky is blue" does not constitute a premise for the assertion "We cannot make brownies", while the sentence "We ran out of chocolate" does because chocolate is an essential ingredient of brownies. However, to classify the relation between the above given premise-claim pair as an instance of reasoning from the formal cause constitutes a task which lies outside common encyclopedic knowledge. In light of this, a set of guidelines about the explicit and implicit components needed to recognize different types of argument schemes between given pairs of premises and claims has been provided.

2.1 The Structure of Argument Schemes following the *Argumentum Model of Topics*

Unlike other contemporary approaches, the *Argumentum Model of Topics* (AMT) does not "conceive of argument schemes as the whole bearing structures that connect the premises to the standpoint or conclusion in a piece of real argumentation" (Rigotti and Morasso, 2010, 483), but as an inference licensed by the combination of both material and procedural premises. Procedural premises are abstract rules of reasoning needed to bridge premises to claims. They include both a broad relation (after which argument schemes are named), which tells us why premises and claims are argumentatively related in a *frame*, and an inferential rule of the implicative type ("if...then"), which further specifies the reasoning at work in drawing a claim from certain premises. Contextual information, necessary to apply abstract rules to a real piece of argumentation, is provided by material premises which include the premise textually expressed and some common ground knowledge about the world. If we consider again the pair of sentences "[We cannot make brownies]CLAIM". "[We ran out of chocolate]PREMISE", the argument scheme connecting them is structured as given in Figure 1. At a structural level, the inferential rule works as a major premise that, combined with the conjunction of the material premises, allows one to draw the conclusion. Among the premises non-textually expressed, while common ground knowledge is per definition accessible to annotators, the inferential rule at work has to be consciously reconstructed.

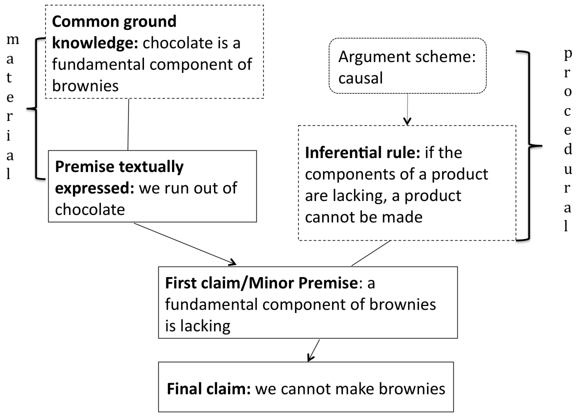
material
Common ground knowledge: chocolate is a fundamental component of brownies
Premise textually expressed: we run out of chocolate
Argument scheme: causal
Inferential rule: if the components of a product are lacking, a product cannot be made
procedural
First claim/Minor Premise: a fundamental component of brownies is lacking
Final claim: we cannot make brownies

Figure 1: Inferential configuration of argument according to *Argumentum Model of Topics (AMT)*

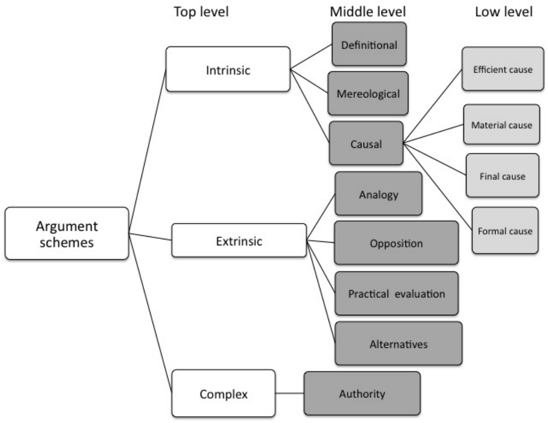

Figure 2: Adopted taxonomy of argument schemes

2.2 A Semantically Motivated Taxonomy of Argument Schemes

In this paper, the adopted taxonomy of argument schemes is a simplified version of that elaborated by exponents of the *Argumentum Model of Topics* (Rigotti, 2006; Palmieri, 2014). According to the AMT, argument schemes are organized in hierarchical clusters based on principles relying on frame semantics and pragmatics. As seen in Figure 2, there are three main levels.

At the top level, argument schemes are distinguished into three groups depending on the type of relations linking the State of Affairs (SoA) expressed by the premise to that expressed by the claim:

- *Intrinsic* argument schemes: the SoA expressed by the premise and that expressed by the claim are linked by an ontological relation since they belong to the same *semantic frame*, understood as a unitarian scene featuring a set of participants (Fillmore and Baker, 2010). This entails that the two SoAs take place simultaneously in the real world or that the existence of one affects the existence of the other.

- *Extrinsic* argument schemes: the SoA expressed by the premise and that expressed by the claim belong to different *frames* and are connected by semantic relations that are not ontological. This means that the existence of one SoA is independent from the existence of the other SoA.

- *Complex* argument schemes: the relation between the SoAs expressed by the premise and the claim is not semantic or ontological, but pragmatic. In other words, what guarantees the support of the claim is reference to an expert or an authority.

The middle level refers to the different types of ontological, semantic and pragmatic relations which further specify the top level classes. Each middle level argument scheme is defined by making reference to semantic or pragmatic properties of the propositions constituting the premises and the conclusion. For example, the scheme *Extrinsic:Practical Evaluation* is defined as follows: "the proposition functioning as premise is an evaluation, namely a judgment about something being 'good' or 'bad'. The claim expresses a recommendation/ advice about stopping/continuing/setting up an action".

The low level further specifies the middle level schemes. For example, the *Intrinsic:Causal* argument scheme is further specified following the so-called Aristotelian causes (efficient cause, formal cause, material cause and final cause)[1]. In the annotation protocol, this low level has not been considered since we hypothesize that it will be difficult for annotators to reliably make such fine-grained distinctions, based on results from similar studies using Walton's taxonomy of argument schemes (Song et al., 2014; Palau and Moens, 2009).

3 Annotation study

The annotation study has been designed on top of the annotation performed by Stab and Gurevych (2014). In their study, annotators were asked to identify and annotate through the open source annotation tool Brat [2] the argumentative components (premise, claim, major claim), the stance characterizing claims (for/against) and the argumentative relations connecting pairs of argumentative components (supports/against) in 90 short persuasive essays. We selected 30 essays as a sample for our pilot annotation (11 relations for each essay in average). The text genre of short persuasive essays is not bound to the discussion of a specific issue, which would prompt the presence of arguments of the same type, but enables the presence of the entire spectrum of argument schemes.

The annotators involved in the project were nine graduate students with no specific background in Linguistics or Argumentation. Three different annotators have been assigned to the annotation of each essay. The task consisted in annotating the "support" relations between premise-claim, claim-major claim, and premise-premise with one of the middle level argumentation schemes given in Figure 2 or *NoArgument*. For the identification of the middle level argument schemes, annotators were provided with an heuristic procedure and asked to look for linguistic clues as a further confirmation for their choices. We included the label of *NoArgument* to account for potential cases where premises/claims in support of claims/major claims do not actually instantiate any inferential path and cannot, hence, be considered proper arguments. For example, in the following pair of clauses: "[This, therefore, makes museums a place to entertain in people leisure time]PREMISE. [People should perceive the value of museums in enhancing their own knowledge]CLAIM ", the clause annotated as premise simply does not underpin at all the clause annotated as claim. As to the "attacks" relations, which indicate that a statement rebuts another statement, they have not been considered as targets of the annotation since they do not directly instantiate an argument scheme linking the spans of texts annotated as premise/claim and claim/major claim, but a complex refutatory move pointing to the defeasibility of the rebutted statement itself or to that of the premises supporting it. Annotators have independently read the guidelines and proceeded with the annotation without any formal training.

The guidelines contain the description of the key notions of argument, premise, claim and argument schemes' components as well as the AMT taxonomy. Detailed instructions about how to proceed in the annotation of argument schemes and rules were provided as well. The main stages of analysis annotators were asked to go through are the following:

- Identification of the middle level argument scheme linking premises-claims or claims-major claims pairs or recognition of the lack of argumentation in doubtful cases (e.g., *Intrinsic:Definitional*, *Intrinsic:Causal*, *Intrinsic:Mereological*, for a total of 9 choices including *NoArgument*, Figure 2)
- Identification of the inferential rule at work

[1] The model presents low level argument schemes for other middle level argument schemes which are not visualized in the Figure 2

[2] http://brat.nlplab.org/

(e.g., Figure 1).

We present these two stages of the annotation process in the next two subsections.

3.1 Identification of the Middle Level Argument Schemes

In order to recognize the middle level types of argument schemes, the annotators were asked to browse a set of given identification questions for argument schemes (see Appendix), to choose the question which best matches the pair of argumentative components linked by a "support" relation, and to check if the argumentative components contain linguistic features listed as typical of an argument scheme (see Appendix).

The explanation of the annotation procedure has been backed up by examples. For instance, given the premise-claim pair: "[Due to the increasing number of petrol stations, the competition in this field is more and more fierce]PREMISE, thus [the cost of petrol could be lower in the future]"CLAIM, the annotators were shown which argument scheme was appropriate:

- *Intrinsic:Definitional*: Does the sentence "Due to the increasing number of petrol stations, the competition in this field is more and more fierce" express a definitional property of the predicate "be lower" attributed to the cost of petrol? NO

  Other linguistic clues: the premise and the claim usually share the grammatical subject. The verb which appears in the claim expresses a state rather than an action.

- *Intrinsic:Mereological*: Is the fact that "Due to the increasing number of petrol stations, the competition in this field is more and more fierce" or an entity of that sentence (e.g., "the competition") an example/a series of examples/a part of the fact that "the cost of petrol could be lower in the future"? NO

  Other linguistic clues: the premise is frequently signaled by the constructions "for example", "as an example", "x proves that".

- *Intrinsic:Causal*: Is the fact that "Due to the increasing number of petrol stations, the competition in this field is more and more fierce" a cause/effect of the fact that "the cost of petrol could be lower in the future" or is it a means to obtain it? YES

  Other linguistic clues: the claim frequently contains a modal verb or a modal construction ("must", "can", "it is clear/it is necessary"). In the given example, the claim contains the modal verb "could".

As far as linguistic clues are concerned, they have been collected from existing literature about linguistic indicators (Rocci, 2012; Miecznikowski and Musi, 2015; Van Eemeren et al., 2007) and from a preliminary analysis of the considered sample. Annotators have been explicitly warned that the given linguistic indicators, due to their highly polysemous and context sensitive nature, do not represent decisive pointers to the presence of specific arguments schemes, but have to be conceived as supplementary measures.

In presence of difficulties to identify a specific argument scheme applying the given set of identification questions, annotators were instructed to embed the pair of argumentative components under the hypothetical construction "If it is true that [premise/claim], is it then true that [claim/major claim]?" and evaluate its soundness. This simple test was meant to help the annotators checking if an inferential relation connecting the argumentative components is possibly there.

If a premise-claim pair failed the test, annotators were asked to choose the label *NoAgument* and explain why argumentation is not there. In the opposite case, they were told to annotate the pair under analysis as *Ambiguous* and try to identify the top level class of argument schemes applying the following round of identification questions:

- *Intrinsic* argument schemes: Can the state of affairs expressed in the premise and the state of affairs expressed in the claim take place simultaneously in the real world or does the realization of one affects the realization of the other one? If yes, it is an instance of intrinsic argument schemes.

- *Extrinsic* argument schemes: Are the existence of the state of affairs expressed in the premise and that expressed in the claim not simultaneous and independent on each other? If yes, it is an instance of extrinsic argument schemes.

- *Complex* argument scheme: Is the premise a discourse/statement expressed by an expert/an authority/an institution and does the claim coincide with the content of that discourse? If yes, it is an instance of complex argument scheme (authority).

Example: Let us consider the example below of a premise supporting a claim.

> "[Knowledge from experience seems a little different from information contained in books]CLAIM . [To cite an example, it is common in books that water boils at 100 Celcius degree. However, the result is not always the same in reality because it also depends on the height, the purity of the water, and even the measuring tool]"PREMISE

To determine whether there is an argument scheme, the annotators could ask themselves: "If it is true that [it is common in books that water boils at 100 Celcius degree. However, the result is not always the same in reality because it also depends on the height, the purity of the water, and even the measuring tool], is it then true that [knowledge from experience seems a little different from information contained in books]?" As the answer is yes, this premise-claim pair is an instance of argument schemes.

When the top level class of argument schemes is concerned, the SoAs expressed by the claim and the premise are simultaneously realized since the premise constitutes an example which shows that what is stated in the claim corresponds to reality. Thus this is an *Intrinsic* scheme. More specifically it is an *Intrinsic: Mereological* scheme (following the questions and the linguistic cues) since a process of induction from an exemplary case to a generalization is at work.

3.2 Identification of the Inferential Rule

The last step of the annotation process consisted in the identification of the inferential rule at work for those pairs in which annotators were able to identify a middle level argument scheme. Annotators were provided with representative rules for each argument scheme (see Appendix) such as the following two for the *Intrinsic:Mereological* argument scheme: "if all parts share a property, then the whole will inherit this property"; "if a part of x has a positive value, also x has a positive value".

They were asked either to write down one of the given inferential rules corresponding to the argument scheme or to formulate a rule on their own if they thought that the provided ones were not fitting. Our hypothesis was that when writing down inferential rules the annotators are forced to control the appropriateness of the chosen argument scheme.

4 Evaluation

In order to evaluate the reliability of the annotations we measured the inter-annotators agreement (IAA) using Fleiss' κ to account for multiple annotators (Fleiss, 1971). When considering the middle level annotation schemes, the IAA is κ=0.1, which shows only slight agreement (Landis and Koch, 1977). This finding supports the hypothesis that for annotating argument schemes the IAA is low when using minimally training non-expert annotators. We also measured the IAA between the top level arguments (*Intrinsic, Extrinsic, Complex, NoArgument*), but did not find any significant difference in the Fleiss' κ score.

Table 1 represents some descriptive statistics about the annotations. Out of 302 argumentative relations to be annotated, for 30 cases (10%) all three annotators agree, while for 179 cases (59%) at least two out of the three annotators agree. When all three annotators agree the distribution of the argument schemes is: 7 *Intrinsic:Causal*, 9 *Intrinsic:Mereorogical*, 1 *Intrinsic:Definitional*, 6 *Extrinsic:Practical Evaluation* and 7 *NoArgument*. When at least two out of the three annotators agree, the distribution of the argument schemes (majority voting) is: 60 (33.5%) *Intrinsic:Causal*, 46 (25.7%) *Intrinsic:Mereorogical*, 16 (8.9%) *Intrinsic:Definitional*, 28 (15.6%) *Extrinsic:Practical Evaluation*, 3(1%) *Extrinsic:Alternatives'* , 3(1%) *Extrinsic:Opposition* and 23 (12.8%) *NoArgument'*).

When considering the 3 top level argument schemes plus *NoArgument*, out of 302 argumentative relations to be annotated, for 260 instances (86%) at least two annotators agreed. The distribution of majority voting labels in these cases is: 185 (71%) are *Intrinsic*, 52 (20%) are *Extrinsic*, and 23 (8.8%) are *NoArgument*.

One goal of this pilot study was to determine whether confusion exists among particular argument schemes with the aim to improve the guidelines. Table 2 shows the confusion matrix between two argument schemes for all annotators pairs. This confusion matrix is a symmetric one, so we

Argument Schemes	# of Agreeing Annotators	# of Instances
Middle	all 3	30
	2 or more	179
Top	all 3	77
	2 or more	260

Table 1: Descriptive Statistics about the annotations

provided only the upper triangular matrix. A detailed discussion is presented in the next section.

5 Discussion of the Results

As shown in the previous section, the argument schemes which received the highest IAA were *Intrinsic:Mereological*, *Intrinsic:Causal* among the *Intrinsic* argument schemes, and *Extrinsic:Practical evaluation* for the *Extrinsic* argument scheme. Going through the examples in which all three annotators agreed, our impression is that both the presence of scheme specific linguistic clues and the suitability of inferential rules already offered in the guidelines enhanced the annotators' choices. As to *Intrinsic:Mereological* relations, the frequent presence of constructions such as "for example", "for instance", compatible only with that specific argument scheme, has plausibly fostered its reliable recognition.

In the case of *Intrinsic:Causal* argument schemes, the cited linguistic clues in the guidelines have turned out to be not relevant: modal verbs are not present in the claims/major claims of the pairs annotated as *Intrinsic: Causal* by the majority of annotators. On the other hand, all these examples are instances of inferential rules from the cause to the effect. This suggests that the cause-effect inferential relation is considered as the prototypical type of causal argument schemes.

Only one instance of *Intrinsic:Definitional* argument scheme was recognized by all three annotators. Notions such as that stative predicates as identifiable linguistic clues in the guidelines were probably not informative for every annotator, as shown by the confusion among the *Intrinsic:Definitional* and *Intrinsic:Causal* argument schemes (Table 2).

For *Intrinsic:Mereological* and *Intrinsic:Causal* argument schemes a set of inferential rules was already proposed in the guidelines, as opposed to just one rule given for *Intrinsic:Definitional.* This has probably helped the annotators to check the soundness of the chosen scheme in these cases.

As to *Extrinsic:Practical Evaluation* argument scheme, the recurrent feature which seems to be at the basis of agreement is the presence of a clear evaluation in the premise.

Table 2 shows that, among the three more frequent argument schemes the *Extrinsic:Practical Evaluation* was the one confused the most with another specific argument scheme, namely *Intrinsic: Causal.* From the analysis of the ambiguous cases, two plausible reasons for the confusion have emerged: 1) the presence of the modal verb "should" has been cited in the guideline among the linguistic clues of both argument schemes, and 2) the *Extrinsic:Practical Evaluation* argument scheme shares with the causal argument scheme of the final type the reference to intentionality and, in general, to the frame of human action where consequences of various choices are taken into account. For example, the premise/claim pair "[this kind of ads will have a negative effect to our children] PREMISE. [Advertising alcohol, cigarettes, goods and services with adult content should be prohibited] CLAIM", which is an instance of *Extrinsic:Practical Evaluation* argument scheme, has been confused with the *Intrinsic:Causal* argument scheme licensing the inferential rule "if an action does not allow to achieve the goal, it should not be undertaken". In order to improve the annotation, ambiguous cases of this type will have to be discussed during the training process (see Section 6).

Since the *Extrinsic:Practical Evaluation* argument scheme is the far most frequent *Extrinsic* argument scheme in our sample, improving its identification promises to highly affect the IAA regarding top level *Extrinsic* vs. *Intrinsic* argument schemes.

The *Intrinsic:Causal* argument scheme appears to be frequently confused also with the *Intrinsic:Mereological* argument schemes and viceversa. This happened mainly in the presence of *Mereological* argument schemes drawing a generalization from a exemplary case (rhetorical induction) such as the following: "[in Vietnam, many cultural costumes and natural scenes, namely drum performance and bay, are being encouraged to preserve and funded by the tourism ministry.] PREMISE [Through tourism industry, many cultural values have been preserved and natural environments have been protected]CLAIM". Some annotators misconceived the SoA expressed by the

	Intrinsic:Causal	I:Mereorogical	I:Definitional	E:PracticalEvaluation	E:Alternatives	E:Opposition	E:Analogy	NoArgument	C:Authority
I:Causal	154	89	45	82	17	17	6	82	5
I:Mereorogical		128	20	47	16	14	14	51	4
I:Definitional			36	26	8	6	5	25	0
E:Practical Evaluation				80	8	8	5	33	1
E:Alternatives					6	3	1	17	0
E:Opposition						14	0	13	1
E:Analogy							0	4	0
NoArgument								74	1
C:Authority									0

Table 2: Confusion Matrix on 30 essays (3 minimally trained non-expert annotators)

premise as an effect of the SoA expressed by the claim. This behavior suggests that the distinction between propositions expressing generalizations and those expressing state of affairs which can be located in space and time was not clear enough in the guidelines.

As to the label *No Argument*, a qualitative analysis of the occurrences showing disagreement has revealed that annotators tried by default to identify an argumentation scheme even when there was none, unless the propositional content of the connected argumentative components was evidently unrelated.

6 Annotation with trained annotators

After the initial study, we improved the guidelines keeping only scheme-specific linguistic clues, providing more inferential rules for each argument scheme, stressing the distinction between *Extrinsic:Practical Evaluation* and *Intrinsic:Causal* as well as between *Intrinsic:Mereological* and *Intrinsic:Causal*, and explicitly stating that some "supports" relations in the corpus are not argumentative (some examples have been provided). In order to test the improvement of the guidelines we have performed a further annotation with 2 trained non-expert annotators and 1 expert annotator on the set of essays which received lowest agreement (κ=-0.01; which indicated poor agreement).

The non-expert annotators went through a two hour training session during which they were asked to annotate 2 essays and received continuous feedback on misunderstandings and/or doubts. The results of the annotation show a shift of the IAA from κ=-0.01 to κ=0.311 ("fair agreement") among all three annotators (including the expert). The IAA among just the two non-expert annotators was similar κ=0.307. In order to map the disagreement space we have calculated the confusion matrix.

Table 3 shows that in this reduced sample the percentage of relations annotated as *No Argument* is higher compared to the overall sample. Looking at the notes made by the annotators, four main reasons for the non argumentative nature of the relations pop up.

First, among the claims-major claims pairs frequently the propositional content of the claim rephrases that of the major claim, such as in the pair "[There should not be any restriction on artists' work]CLAIM. [The artist must be given freedom]"MAJOR CLAIM. In these cases, the presence of a "supports" relation is justified if redundancy is considered as a stylistic strategy for achieving consensus on a certain stance; however, the claim as a linguistic entity does not work as a argument.

Second, the clause annotated as premise happened to work as an argument only if combined with another clause. This happens bacause the annotation of premises and claims in the original dataset of Stab and Gurevych (2014) was done at the clause level. As recently pointed out by Stede et al. (2016) the mismatch between ADUs, which tend to encompass multiple clauses, and EDUs (elementary discourse units), constitutes one of the major difficulties to overcome in the investigation of the existing intersections between argumentative and discourse relations.

Third, the relation between two argumentative components would have been argumentative if reversed, or if a different claim would have been

	Intrinsic:Causal	I:Mereorogical	I:Definitional	E:PracticalEvaluation	E:Alternatives	E:Opposition	E:Analogy	NoArgument	C:Authority
I:Causal	86	19	10	13	0	1	0	47	0
I:Mereorogical		70	5	1	0	0	0	21	0
I:Definitional			0	1	0	0	0	10	0
E:PracticalEvaluation				10	0	0	0	9	0
E:Alternatives					0	0	0	0	0
E:Opposition						2	0	4	0
E:Analogy							0	0	0
No Argument								136	0
C:Authority									0

Table 3: Confusion Matrix on a set of 10 essays (highly trained annotators: 2 non-experts and 1 expert)

chosen.

Fourth, the clause annotated as premise does not underpin in anyway the clause annotated as claim, but constitutes instead a counterargument.

Although the agreement in the recognition of *No Argument* cases has consistently improved with highly trained annotators (non-expert as well as expert), it still remains a matter of confusion. In particular, the most frequent label chosen instead of *NoArgument* is that of *Intrinsic:Causal* argument scheme. This is probably due to the implicative nature of the proposed test "if the premise is true, then the claim is true", which invites a causal interpretation.

7 Conclusion and Future Work

We presented a novel set of guidelines for the annotation of argument schemes based on the *Argumentum Model of Topics*. This framework is advantageous since it offers a hierarchical finite taxonomy of argument schemes based on linguistic criteria which are highly distinctive and applicable to every context. We have conducted a pilot annotation study of 30 short persuasive essays with 9 minimally trained non-expert annotators in order to test the informativeness of the guidelines. The low inter-annotator agreement confirms the difficulties underlined by previous studies for minimally trained annotators to recognize argument schemes. From the qualitative analysis of the confusion matrixes it has emerged that: 1) linguistic indicators of argument schemes constitute useful clues for the annotators only if specific to one argument scheme, otherwise they can be a source of confusion; 2) the reconstruction of inferential rules is highly relevant to enhancing annotators' choices and 3) among *Intrinsic:Causal* argument schemes the subtype "Efficient cause" is the easiest to identify. We have improved the guidelines according to these results and tested them on a reduced sample of 10 essays with 2 trained non-expert annotators and one expert annotator. The interannotator agreement has significantly improved (fair agreement). The confusion matrix suggests that the frequency of non argumentative or ambiguous relations is the main cause of disagreement. For future work, we plan to test again the annotation guidelines in a corpus with higher accuracy as to the annotation of argumentative components (premises/claims). A methodological result of the study is that identifying argument schemes constitutes an important tool to verify the presence of argumentative components, and support relations.

Acknowledgements

This paper is based on work supported partially by the Early Post Doc SNFS Grant for the project "From semantics to argumentation mining in context: the role of evidential strategies as indicators of argumentative discourse relations" and DARPA-DEFT program. The views expressed are those of the authors and do not reflect the official policy or position of the SNFS, Department of Defense or the U.S. Government. We would like to thank the annotators for their work and the anonymous reviewers for their valuable feedback.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfre-

und, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 162–168. IEEE.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Charles J Fillmore and Collin Baker. 2010. A frames approach to semantic analysis. *The Oxford handbook of linguistic analysis*, pages 313–339.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.

Nancy L Green. 2015. Identifying argumentation schemes in genetics research articles. *NAACL HLT 2015*, page 12.

Scott Jacobs. 2000. Rhetoric and dialectic from the standpoint of normative pragmatics. *Argumentation*, 14(3):261–286.

Joel Katzav and Chris A Reed. 2004. On argumentation schemes and the natural classification of arguments. *Argumentation*, 18(2):239–259.

Manfred Kienpointner. 1987. Towards a typology of argumentative schemes. *Argumentation: Across the lines of discipline*, 3:275–87.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *LREC*, pages 462–468.

Fabrizio Macagno. 2014. Argumentation schemes and topical relations. *Macagno, F. & Walton, D.(2014). Argumentation schemes and topical relations. In G. Gobber, and A. Rocci (eds.), Language, reason and education*, pages 185–216.

Johanna Miecznikowski and Elena Musi. 2015. Verbs of appearance and argument schemes: Italian sembrare as an argumentative indicator. In *Reflections on Theoretical Issues in Argumentation Theory*, pages 259–278. Springer.

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

Rudi Palmieri. 2014. *Corporate argumentation in takeover bids*, volume 8. John Benjamins Publishing Company.

Andreas Peldszus and Manfred Stede. 2013a. 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. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 196–204.

John L Pollock. 1995. *Cognitive carpentry: A blueprint for how to build a person*. Mit Press.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Eddo Rigotti and Sara Greco Morasso. 2010. Comparing the argumentum model of topics to other contemporary approaches to argument schemes: the procedural and material components. *Argumentation*, 24(4):489–512.

Eddo Rigotti. 2006. Relevance of context-bound loci to topical potential in the argumentation stage. *Argumentation*, 20(4):519–540.

Eddo Rigotti. 2009. Whether and how classical topics can be revived within contemporary argumentation theory. In *Pondering on problems of argumentation*, pages 157–178. Springer.

Andrea Rocci. 2012. Modality and argumentative discourse relations: a study of the italian necessity modal dovere. *Journal of Pragmatics*, 44(15):2129–2149.

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

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1069–1080. ACM.

Yi Song, Michael Heilman, Beata Beigman, and Klebanov Paul Deane. 2014. Applying argumentation schemes for essay scoring.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémie Perret. 2016. Parallel discourse annotations on a corpus of short texts.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Frans H Van Eemeren and Rob Grootendorst. 1992. *Argumentation, communication, and fallacies: A pragma-dialectical perspective.* Lawrence Erlbaum Associates, Inc.

Frans H Van Eemeren, Peter Houtlosser, and AF Snoeck Henkemans. 2007. *Argumentative indicators in discourse: A pragma-dialectical study*, volume 12. Springer Science & Business Media.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments.* Routledge.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes.* Cambridge University Press.

Douglas Walton. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(3):33–61.

A Appendix

We report in what follows the "cheat sheet" located at the end of the annotation guidelines which contains i) an identification question, ii) a set of linguistic clues and of iii) inferential relations for each middle level argument scheme. The complete guidelines will be made available.

1. Intrinsic Definition:

   Does x express a definitional property of the predicate attributed to the grammatical subject in y?

   Other clues: the premise and the claim usually share the grammatical subject. The verb which appears in the claim expresses a state (*be +noun or be + adjective, consider*) rather than an action.

   Inferential rule: "if x shows typical traits of a class of entities (e.g. positive actions, beneficial decisions), then it is an instance of that class"

2. Intrinsic Mereorogical:

   Is "the fact that x" or an entity cited in x an example /a series of examples /a part of "the fact that y"?

   Other clues: the premise is frequently signaled by the constructions *or example, as an example, for instance, x proves that.* In the cases in which induction is at work the premise coincides with the description of a situation that is frequently located in the past.

   Inferential rules:

   - "if all parts share property, then the whole will inherit this property"
   - "if a part of x has a positive value, also x has a positive value"
   - "if something holds/may hold/held for an exemplary case x, it holds/may hold/will hold for all the cases of the same type"
   - "if something holds/may hold/held for a sample of cases of the type x, it holds/may hold/will hold for every case of the type x"

3. Intrinsic Causal:

   Is x a cause /effect of y or is it a means to obtain y?

   Other clues: the claim frequently contains a modal verb or a modal construction (*must, can, it is clear /it is necessary*).

   Inferential rules:

   - "if the cause is the case, the effect is the case"
   - "if the effect is the case, the cause is probably the case"
   - "if a quality characterizes the cause, then such quality characterizes the effect too"

- "if the realization of the goal necessitates the means x, x must be adopted"
- "if an action does not allow to achieve the goal, it should not be undertaken"
- "if somebody has the means to achieve a certain goal, he will achieve that goal"

4. Extrinsic Analogy:

   Do x and/or y compare situations happened in different circumstances but similar in some respects?

   Other clues: the premise and /or the claim usually contain comparative conjunctions /constructions (e.g. *as, like, in a similar vein*)

   Inferential rules:

   - "if the state of affairs x shows a set of features which are also present in the state of affairs y and z holds for x, then z holds for y too"
   - "if two events x and y are similar and event x had the consequence z, probably also y will have the consequence z"
   - "if two situations x and y are similar in a substantial way and action z was right in the situation x, action y will be right also in the situation y"

5. Extrinsic Opposition:

   does the occurrence of the state of affairs x exclude the occurrence of the state of affairs y? Or does the premise contain entities /events which are opposite with respect entities /events expressed in the claim?

   Other clues: the claim sometimes contain modals which express impossibility (*it is impossible that, it cannot be that*, but it is not always the case.

   Inferential rules:

   - "If two state of affairs/entities x, y are one the opposite of the other, the occurrence of x excludes the occurrence of y"
   - "If two state of affairs x, y are one the opposite of the other, they entail opposite consequences"

6. Extrinsic Alternatives:

   Is/are the state of affairs expressed by x an alternative(s) to the one expressed in y?

   Other clues: the claim frequently contains necessity modals (*must, have to*). The premise states that all possible other alternatives are excluded.

   Inferential rules:

   - "if all the alternatives to x are excluded, then x is unavoidable"
   - "if among a set of alternatives only one is reasonable it has to be undertaken"

7. Extrinsic Practical Evaluation/Termination and setting up:

   Does x express an evaluation and does y express an /a recommendation about stopping /continuing /setting up that action?

   Other clues: the claim usually contains the modal verb *should.*

   Inferential rules:

   - "if something is of important value, it should not be terminated"
   - "if something has a positive value, it should be supported /continued /promoted /maintained"
   - "if something has positive effects, it should be supported /continued /promoted /maintained"
   - "if something has a negative effect it should be terminated"

8. Complex Authority:

   Is the premise a discourse/statement expressed by a an expert /institution /authority in the field and does the claim coincides with the content of that discourse?

   Other clues: the authority to which the writer appeals is usually introduced by *according to, as shown by, as clarified /explained /declared by*.

   Inferential rules:

   - "if the institution /expert /authority in the field states that proposition x is true, then x is true"
   - "if the institution /expert /authority in the field states event x will occur, then x will probably occur"