HAL
archives-ouvertes.fr

Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study

Elena Cabrio, Serena Villata

▶ To cite this version:

Elena Cabrio, Serena Villata. Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study. Joint Symposium on Semantic Processing (JSSP-2013), Nov 2013, Trento, Italy. hal-00915879

HAL Id: hal-00915879
https://hal.inria.fr/hal-00915879

Submitted on 26 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study.

Elena Cabrio and **Serena Villata**
INRIA
2004 Route des Lucioles BP93
06902 Sophia-Antipolis cedex, France.
{elena.cabrio, serena.villata}@inria.fr

Abstract

In the knowledge representation and reasoning research area, argumentation theory aims at representing and reasoning over information items called *arguments*. In everyday life, arguments are reasons to believe and reasons to act, and they are usually expressed in natural language. Even if *ad-hoc* natural language examples are often provided in argumentation theory works, no automated processing of such natural language arguments is carried out, making it impossible to exploit the results of this research area in real world scenarios. In this paper, we propose to adopt textual entailment to address this issue. In particular, we discuss and evaluate, on a sample of natural language arguments extracted from Debatepedia, the support and attack relations among arguments in bipolar abstract argumentation with respect to the more specific notions of textual entailment and contradiction.

1 Introduction

Until recent years, the idea of "argumentation" as the process of creating arguments for and against competing claims was a subject of interest to philosophers and lawyers. In recent years, however, there has been a growth of interest in the subject from formal and technical perspectives in Artificial Intelligence, and a wide use of argumentation technologies in practical applications. However, such applications are always constrained by the fact that natural language arguments cannot be automatically processed by such argumentation technologies. Arguments are usually presented either as the abstract nodes of a directed graph where the edges represent the relations of attack and support (e.g., in abstract argumentation theory (Dung, 1995) and in bipolar argumentation (Cayrol and Lagasquie-Schiex, 2005)) or as a set of premises which lead to a certain conclusion thanks to the application of a number of inference rules (e.g., in structured approaches to argumentation as ASPIC (Prakken, 2010)). Natural language arguments are usually used in such works to provide *ad-hoc* examples to help the reader in the understanding of the rationale behind the formal approach which is then introduced, but the need to find automatic ways to process natural language arguments to detect the semantic relations among them is becoming more and more important.

To fill this gap, we propose to investigate semantic inference approaches in Natural Language Processing (NLP) in search of a suitable computational framework to account for bipolar argumentation models. In particular, in this paper, we study *how bipolar semantic relations among natural language arguments can be discovered in an automated way using textual entailment.* This issue breaks down into the following research questions: (1) what is the relation between the notion of support in bipolar argumentation and the notion of Textual Entailment (TE) in NLP?, and given that additional attacks have been proposed in the literature to highlight possible inconsistencies arising among sets of arguments connected by supports and attacks (2) what is the distribution and thus the inner semantics of such additional attacks in real data?

First, we study the relation among the notion of support in bipolar argumentation (Cayrol and Lagasquie-Schiex, 2005), and the notion of TE in NLP (Dagan et al., 2009). In a recent proposal, (Cabrio and Villata, 2012) represent the TE relation extracted from NL texts as a support relation in bipolar argumentation. This is a strong assumption, and we aim at verifying on a sample of real data from Debatepedia whether it is always the case that support is equivalent to TE. In partic-

ular, for addressing this issue, we focus both on the relations between support and entailment, and on the relations between attack and contradiction. We show that TE and contradiction are more specific concepts than support and attack, but still hold in most of the argument pairs.

Second, starting from the comparative study addressed by (Cayrol and Lagasquie-Schiex, 2011), we consider four additional attacks proposed in the literature: *supported* (if argument a supports argument b and b attacks argument c, then a attacks c) and *secondary* (if a supports b and c attacks a, then c attacks b) attacks (Cayrol and Lagasquie-Schiex, 2010), *mediated* attacks (Boella et al., 2010) (if a supports b and c attacks b, then c attacks a), and *extended* attacks (Nouioua and Risch, 2010; Nouioua and Risch, 2011) (if a supports b and a attacks c, then b attacks c). We investigate the distribution of these attacks in NL debates basing on a data set extracted from Debatepedia, and we show that all these models are verified in human debates, even if with a different frequency.

The benefit of the proposed analysis is twofold. First, it is used to verify, through a data driven evaluation, the "goodness" of the proposed models of bipolar argumentation to be used in real settings, going beyond *ad hoc* NL examples. Second, it can be used to guide the construction of cognitive agents whose major need is to achieve a behavior as close as possible to the human one. Thanks to such a kind of analysis, we highlight that the mutual influence of these two related research areas can actually bring to textual entailment more than just an application scenario, but it opens further challenges to be addressed with a joint effort by the two research communities.

The paper is organized as follows. Section 2 summarizes the basic notions of bipolar argumentation, and describes the four kinds of additional attacks we consider in this paper. Section 3 describes the experimental setting, and addresses the analysis of the meaning of support and attack in natural language dialogues, as well as the comparative study on the existing additional attacks.

2 Bipolar argumentation

We provide the basic concepts of Dung's (1995) abstract argumentation.

Definition 1 *(Abstract argumentation framework) An abstract argumentation framework (AF) is a pair $\langle A, \rightarrow \rangle$ where A is a set of elements called arguments and $\rightarrow \subseteq A \times A$ is a binary relation called attack. We say that an argument a attacks an argument b if and only if $(a, b) \in \rightarrow$.*

Dung presents several acceptability semantics that produce zero, one, or several sets of accepted arguments called *extensions*. For more details, see (Dung, 1995).

Bipolar argumentation frameworks, firstly proposed by (Cayrol and Lagasquie-Schiex, 2005), extend Dung's framework taking into account both the attack relation and the support relation. In particular, an abstract bipolar argumentation framework is a labeled directed graph, with two labels indicating either attack or support. In this paper, we represent the attack relation by $a \rightarrow b$, and the support relation by $a \dashrightarrow b$.

Definition 2 (Bipolar argumentation framework) *A bipolar argumentation framework (BAF) is a tuple $\langle A, \rightarrow, \dashrightarrow \rangle$ where A is the set of elements called arguments, and two binary relations over A are called attack and support, respectively.*

(Cayrol and Lagasquie-Schiex, 2011) address a formal analysis of the models of support in bipolar argumentation to achieve a better understanding of this notion and its uses. In the rest of the paper, we will adopt their terminology to refer to additional attacks, i.e., *complex attacks*. (Cayrol and Lagasquie-Schiex, 2005; Cayrol and Lagasquie-Schiex, 2010) argue about the emergence of new kinds of attacks from the interaction between the attacks and supports in BAF. In particular, they specify two kinds of complex attacks called *secondary* and *supported* attacks, respectively.

Definition 3 (Secondary and supported attacks) *Let BAF = $\langle A, \rightarrow, \dashrightarrow \rangle$ where $a, b \in A$. A supported attack for b by a is a sequence $a_1 R_1 \ldots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $\forall i = 1 \ldots n-2, R_i = \dashrightarrow$ and $R_{n-1} = \rightarrow$. A secondary attack for b by a is a sequence $a_1 R_1 \ldots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $R_1 = \rightarrow$ and $\forall i = 2 \ldots n\text{-}1, R_i = \dashrightarrow$.*

According to the above definition, these attacks hold in the first two cases depicted in Figure 1, where there is a supported attack from a to c, and there is a secondary attack from c to b.

The support relation has been specialized in other approaches where new complex attacks emerging from the combination of existing attacks and supports are proposed. (Boella et al., 2010)

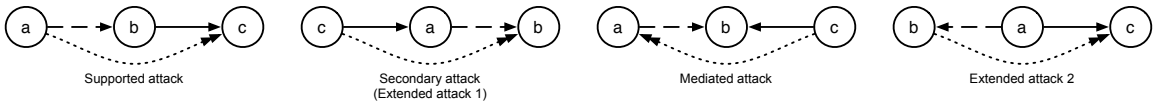

Figure 1: Additional attacks emerging from the interaction of supports and attacks.

propose a *deductive* view of support in abstract argumentation where, given the support $a \dashrightarrow b$ the acceptance of a implies the acceptance of b, and the rejection of b implies the rejection of a. They introduce a new kind of complex attack called *mediated* attacks (Figure 1).

Definition 4 (Mediated attacks) *Let BAF* $= \langle A, \rightarrow, \dashrightarrow \rangle$ *where* $a, b \in A$. *A mediated attack on* b *by* a *is a sequence* $a_1 R_1 \ldots R_{n-2} a_{n-1}$ *and* $a_n R_{n-1} a_{n-1}$, $n \geq 3$, *with* $a_1 = a, a_{n-1} = b, a_n = c$, *such that* $R_{n-1} = \rightarrow$ *and* $\forall i = 1 \ldots n\text{-}2, R_i = \dashrightarrow$.

(Nouioua and Risch, 2010; Nouioua and Risch, 2011) propose, instead, an account of support called *necessary* support. In this framework, given $a \dashrightarrow b$ then the acceptance of a is necessary to get the acceptance of b, i.e., the acceptance of b implies the acceptance of a. They introduce two new kinds of complex attacks called *extended attacks* (Figure 1). Note that the first kind of extended attacks is equivalent to the secondary attacks introduced by (Cayrol and Lagasquie-Schiex, 2005; Cayrol and Lagasquie-Schiex, 2010), and that the second case is the dual of supported attacks.

Definition 5 (Extended attacks) *Let BAF* $= \langle A, \rightarrow, \dashrightarrow \rangle$ *where* $a, b \in A$. *An extended attack on* b *by* a *is a sequence* $a_1 R_1 a_2 R_2 \ldots R_n a_n$, $n \geq 3$, *with* $a_1 = a, a_n = b$, *such that* $R_1 = \rightarrow$ *and* $\forall i = 2 \ldots n, R_i = \dashrightarrow$, *or a sequence* $a_1 R_1 \ldots R_n a_n$ *and* $a_1 R_p a_p$, $n \geq 2$, *with* $a_n = a, a_p = b$, *such that* $R_p = \rightarrow$ *and* $\forall i = 1 \ldots n, R_i = \dashrightarrow$.

All these models of support in bipolar argumentation address the problem of how computing the set of extensions from the extended framework providing different kinds of solutions, i.e., introducing the notion of *safety* in BAF (Cayrol and Lagasquie-Schiex, 2005), or computing the extensions in the meta-level (Boella et al., 2010; Cayrol and Lagasquie-Schiex, 2010). In this paper, we are not interested in discussing and evaluating these different solutions. Our aim is to evaluate how much these different models of support occur and are effectively "exploited" in NL dialogues, to provide a better understanding of the notion of support and attack in bipolar argumentation.

We are aware that the notion of support is a controversial one in the field of argumentation theory. In particular, another view of support sees this relation as a relation holding among the premises and the conclusion of a structured argument, and not as another relation among the arguments (Prakken, 2010). However, given the amount of attention bipolar argumentation is receiving in the literature (Rahwan and Simari, 2009), a better account of this kind of frameworks is required.

Another approach to support has been proposed by (Oren and Norman, 2008; Oren et al., 2010) where they distinguish among *prima-facie* arguments and standard ones. They show how a set of arguments described using Dung's argumentation framework can be mapped from and to an argumentation framework that includes both attack and support relations. The idea is that an argument can be accepted only if there is an evidence supporting it, i.e., evidence is represented by means of *prima-facie* arguments. In this paper, we concentrate our analysis on the abstract models of bipolar argumentation proposed in the literature (Cayrol and Lagasquie-Schiex, 2010; Boella et al., 2010; Nouioua and Risch, 2011), and we leave as future work the account of support in structured argumentation and the model proposed by (Oren and Norman, 2008; Oren et al., 2010).

3 Empirical studies on NL debates

Starting from (Cabrio and Villata, 2012), as a case study to carry out our analysis we select Debatepedia[1], the Wikipedia of debates. Specifically, Debatepedia is an encyclopedia of *pro* and *con* arguments where users can freely contribute to online discussions about critical issues. We collect a sample of the discussions extracting a set of arguments from Debatepedia topics, as described in Section 3.1. Even if our data set cannot be exhaustive, the methodology we apply for the arguments extraction aims at preserving the original structure

[1] http://idebate.org

of the debate, to make it as representative as possible of human daily natural language interactions.

Two different empirical studies are then presented in this section. The first one (Section 3.2) starts from (Cabrio and Villata, 2012), and explores the relation among the notion of *support* and *attack* in bipolar argumentation, and the *semantic inferences* as defined in the NLP research field. The second analysis (Section 3.3) starts instead from the comparative study of (Cayrol and Lagasquie-Schiex, 2011) of the four complex attacks proposed in the literature, and investigates their distribution in NL debates.

3.1 Data set

To have a stable version of the data to perform our studies, we build a reference data set extracting a sample of debates from Debatepedia[2]. Here, the users manually insert their arguments in the column PRO if they agree with the issue under discussion, or in the column CON if they disagree. To make our sample of NL debates comparable with current works in the literature, e.g. (Wyner and van Engers, 2010; Carenini and Moore, 2006; Cabrio and Villata, 2012), we select the same topics as (Cabrio and Villata, 2012), since this is the only freely available data set of natural language arguments (Table 1, column *Topics*). To create the Debatepedia data set, for each topic of our sample we apply the following procedure:

1. the main issue (i.e., the title of the debate in its affirmative form) is considered as the starting argument;
2. each user opinion is extracted and considered as an argument;
3. since *attack* and *support* are binary relations, the arguments are coupled with:
   (a) the starting argument, or
   (b) other arguments in the same discussion to which the most recent argument refers (e.g., when a user opinion supports or attacks an argument previously expressed by another user), following the chronological order (we maintain the dialogue structure);
4. the resulting pairs of arguments are then tagged with the appropriate relation, i.e., *attack* or *support*.

[2] http://bit.ly/VZIs6M

To show a step-by-step application of the procedure, let us consider the debated issue *Can coca be classified as a narcotic?*. At step 1, we transform its title into the affirmative form, and we consider it as the starting argument **(a)**:

(a) *Coca can be classified as a narcotic.*

At step 2, we extract all the users opinions on this issue (PRO and CON), e.g., **(b)**, **(c)** and **(d)**:

(b) *In 1992 the World Health Organization's Expert Committee on Drug Dependence (ECDD) undertook a 'prereview' of coca leaf at its 28th meeting. The 28th ECDD report concluded that, "the coca leaf is appropriately scheduled as a narcotic under the Single Convention on Narcotic Drugs, 1961, since cocaine is readily extractable from the leaf." This ease of extraction makes coca and cocaine inextricably linked. Therefore, because cocaine is defined as a narcotic, coca must also be defined in this way.*

(c) *Coca in its natural state is not a narcotic. What is absurd about the 1961 convention is that it considers the coca leaf in its natural, unaltered state to be a narcotic. The paste or the concentrate that is extracted from the coca leaf, commonly known as cocaine, is indeed a narcotic, but the plant itself is not.*

(d) *Coca is not cocaine. Coca is distinct from cocaine. Coca is a natural leaf with very mild effects when chewed. Cocaine is a highly processed and concentrated drug using derivatives from coca, and therefore should not be considered as a narcotic.*

At step 3a we couple the arguments **(b)** and **(d)** with the starting issue since they are directly linked with it, and at step 3b we couple argument **(c)** with argument **(b)**, and arguments **(d)** with argument **(c)** since they follow one another in the discussion. At step 4, the resulting pairs of arguments are then tagged with the appropriate relation: **(b)** *supports* **(a)**, **(d)** *attacks* **(a)**, **(c)** *attacks* **(b)** and **(d)** *supports* **(c)**.

Table 1 reports the number of arguments and pairs we extracted applying the extraction methodology described before to all the mentioned topics. In total, our data set contains 310 different arguments and 320 argument pairs (179 expressing the *support* relation among the involved arguments, and 141 expressing the *attack* relation). We consider the obtained data set as representative

of human debates in a non-controlled setting (Debatepedia users position their arguments w.r.t. the others as PRO or CON, the data are not biased), and therefore we use it for our empirical studies.

DEBATEPEDIA data set		
Topic	**#argum**	**#pairs**
VIOLENT GAMES BOOST AGGRESSIVENESS	17	23
CHINA ONE-CHILD POLICY	11	14
CONSIDER COCA AS A NARCOTIC	17	22
CHILD BEAUTY CONTESTS	13	17
ARMING LIBYAN REBELS	13	15
RANDOM ALCOHOL BREATH TESTS	11	14
OSAMA DEATH PHOTO	22	24
PRIVATIZING SOCIAL SECURITY	12	13
INTERNET ACCESS AS A RIGHT	15	17
GROUND ZERO MOSQUE	11	12
MANDATORY MILITARY SERVICE	15	17
NO FLY ZONE OVER LIBYA	18	19
AIRPORT SECURITY PROFILING	12	13
SOLAR ENERGY	18	19
NATURAL GAS VEHICLES	16	17
USE OF CELL PHONES WHILE DRIVING	16	16
MARIJUANA LEGALIZATION	23	25
GAY MARRIAGE AS A RIGHT	10	10
VEGETARIANISM	14	13
TOTAL	310	**320**

Table 1: Debatepedia data set.

3.2 First study: support and TE

Our first empirical study aims at a better understanding of the relation among the notion of support in bipolar argumentation (Cayrol and Lagasquie-Schiex, 2011), and the definition of semantic inference in NLP (in particular, the more specific notion of TE) (Dagan et al., 2009). In a recent work, (Cabrio and Villata, 2012) propose to combine NLP and Dung-like abstract argumentation to generate the arguments from NL text, and compute the accepted ones. They represent the TE relation as a support relation in BAF. Even if they narrow their work by considering only favorable arguments implying another argument, explicitly stating that arguments supporting another argument but without inferring it are out of the scope of that work, the assumption that there exists an identity between support and TE is still a claim to verify.

3.2.1 Textual Entailment

The notion of TE has been defined as a directional relation between two textual fragments, termed *text (T)* and *hypothesis (H)*, respectively (Dagan et al., 2009). The relation holds (i.e. $T \Rightarrow H$) whenever the truth of one text fragment follows from the other, as interpreted by a typical language user. Let us consider for instance the two textual fragments **(a)** and **(b)** from Debatepedia. According to the TE framework we set **(b)** as T and **(a)** as H:

(b) $\mapsto$ **T:** *In 1992 the World Health Organization's Expert Committee on Drug Dependence (ECDD) undertook a 'pre-review' of coca leaf at its 28th meeting. [. . .] This ease of extraction makes coca and cocaine inextricably linked. Therefore, because cocaine is defined as a narcotic, coca must also be defined in this way.*
(a)$\mapsto$ **H:** *Coca can be classified as a narcotic.*

A human reading T would infer that H is most likely true (i.e. the meaning of H can be derived from the meaning of T, so the entailment holds). On the contrary, if we consider Debatepedia examples **(a)** and **(d)**, and we set **(d)** as T and **(a)** as H, there is a contradiction between T and H:

(d) $\mapsto$ **T:** *Coca is not cocaine. Coca is distinct from cocaine. Coca is a natural leaf with very mild effects when chewed. Cocaine is a highly processed and concentrated drug using derivatives from coca, and therefore should not be considered as a narcotic.*
(a)$\mapsto$ **H:** *Coca can be classified as a narcotic.*

(de Marneffe et al., 2008) provide a definition of contradiction for the TE task, claiming that it occurs when two sentences *i)* are extremely unlikely to be true simultaneously, and *ii)* involve the same event. As an applied framework, TE has been proposed to capture major semantic inference needs across NLP applications (e.g., question answering, information extraction).

3.2.2 Analysis on the Debatepedia data set

Based on the TE definition, an annotator with skills in linguistics has carried out a first phase of annotation of the Debatepedia data set (Section 3.1). The goal of such annotation is to individually consider each pair of *support* and *attack* among arguments, and to additionally tag them as *entailment*, *contradiction* or *null*. The *null* judgment can be assigned in case an argument is supporting another argument without inferring it, or the argument is attacking another argument without contradicting it. As exemplified above, a correct entailment pair is **(b) entails (a)**, while a contradiction is **(d) contradicts (a)**. A *null* judgment is assigned to **(d) - (c)**, since the former argument supports the latter without inferring it. Our data set is an extended version of (Cabrio and Villata, 2012)'s one allowing for a deeper investigation.

To assess the validity of the annotation task, we

calculated the inter-annotator agreement. Another annotator with skills in linguistics has therefore independently annotated a sample of 100 pairs of the data set. To calculate the inter-rater agreement we used Cohen's kappa coefficient (Carletta, 1996). For NLP tasks, the agreement is considered as significant when κ >0.6. We calculated the inter-annotator agreement on the argument pairs tagged as *support* and *attacks* by both annotators. For supports, we calculated the agreement between the pairs tagged as *entailment* and as *null* (i.e. no entailment); for the contradictions, the agreement between the pairs tagged as *contradiction* and as *null* (i.e. no contradiction). Applying κ to our data, the agreement for our task is $\kappa = 0.74$, that is a satisfactory agreement.

Table 2 reports the results of the annotation on our Debatepedia data set, as resulting after a reconciliation phase carried out by the annotators.

Relations		%arguments (#arg)
support	+ *entailment*	61.6 (111)
	- *entailment (null)*	38.4 (69)
attack	+ *contradiction*	71.4 (100)
	- *contradiction (null)*	28.6 (40)

Table 2: Support and TE on Debatepedia data set.

On the 320 pairs of the data set, 180 represent a *support* relation, while 140 are *attacks*. Considering only the *supports*, we can see that 111 argument pairs (i.e., 61.6%) are an actual entailment, while in 38.4% of the cases the first argument of the pair supports the second one without inferring it (as for example **(d) - (c)**). With respect to the *attacks*, we can notice that 100 argument pairs (i.e., 71.4%) are both attack and contradiction, while only the 28.6% of the argument pairs does not contradict the arguments they are attacking, as in the following example:

(e) *Coca chewing is bad for human health. The decision to ban coca chewing fifty years ago was based on a 1950 report elaborated by the UN Commission of Inquiry on the Coca Leaf with a mandate from ECOSOC: "We believe that the daily, inveterate use of coca leaves by chewing is thoroughly noxious and therefore detrimental".*

(f) *Chewing coca offers an energy boost. Coca provides an energy boost for working or for combating fatigue and cold.*

Differently from the relationship between support-entailment, the difference between attack and contradiction is more subtle, and it is not always straightforward to say when an argument attacks another argument without contradicting it. In the example, we consider that **(e)** does not explicitly contradict **(f)** even if it attacks **(f)**, since chewing coca can offer an energy boost, and still be bad for human health. As we can notice from the results in Table 2, this kind of attacks is less frequent than the attacks-contradictions.

Considering the TE three way scenario (*entailment*, *contradiction*, *unknown*) to map TE relation with bipolar argumentation, argument pairs connected by a relation of support (but where the first argument does not entail the second one), and argument pairs connected by a relation of attack (but where the first argument does not contradict the second one) have to be mapped as *unknown* pairs in the TE framework. The *unknown* relation in TE refers to the T-H pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T. This is a broad definition, that can apply also to pairs of non related sentences (that are considered as unrelated arguments in bipolar argumentation).

From an application viewpoint, as highlighted in (Reed and Grasso, 2007; Heras et al., 2010), argumentation theory should be used as a tool in online discussions applications to identify the relations among the statements, and to provide a structure to the dialogue to easily evaluate the user's opinions. Starting from the methodology proposed by (Cabrio and Villata, 2012) for passing from NL arguments to a Dung's system towards a fully automated system to identify the accepted arguments, our study demonstrates that applying the TE approach would be productive in the 66% of the Debatepedia data set. Other techniques should be investigated to cover the other cases, for instance measuring the semantic relatedness of the two propositions, e.g., Latent Semantics Analysis techniques (Landauer et al., 1997).

3.3 Second study: complex attacks

As a second step of our survey, we carry out a comparative evaluation of the four proposals of attacks suggested in the literature, and we investigate their distribution and meaning on the sample of NL arguments.

3.3.1 Analysis on the Debatepedia data set

Relying on the additional attacks (Section 2), and the original AF of each topic in our data set (Table 1), the following procedure is applied: the *supported* (secondary, mediated, and extended, re-

spectively) attacks are added, and the argument pairs resulting from coupling the arguments linked by this relation are collected in the data set "supported (secondary, mediated, and extended, respectively) attack".

Collecting the arguments pairs generated from the different types of complex attacks in separate data sets allows us to independently analyze each type, and to perform a more accurate evaluation.[3] Figures 2a-d show the four AFs resulting from the addition of the complex attacks in the example *Can coca be classified as a narcotic?*. The reader may observe that the AF in Figure 2a, where the supported attack is introduced, is the same of Figure 2b where the mediated attack is introduced. Notice that, even if the attack which is introduced is the same, i.e., d attacks b, this is due to different interactions among supports and attacks (as highlighted in the figure), i.e., in the case of supported attacks this is due to the support from d to c and the attack from c to b, while in the case of mediated attacks this is due to the support from b to a and the attack from d to a.

A second annotation phase is then carried out on the data set, to verify if the generated arguments pairs of the four data sets are actually attacks (i.e., if the models of complex attacks proposed in the literature are represented in real data). More specifically, an arguments pair resulting from the application of a complex attack can be annotated as: *attack* (if it is a correct attack) or as *unrelated* (in case the meanings of the two arguments are not in conflict). For instance, the pair **(g)-(h)** resulting from the insertion of a *supported* attack, cannot be considered as an attack since the arguments are considering two different aspects of the issue.

(g) *Chewing coca offers an energy boost. Coca provides an energy boost for working or for combating fatigue and cold.*
(h) *Coca can be classified as a narcotic.*

In the annotation, *attacks* are then annotated also as *contradiction* (if the first argument contradicts the other) or *null* (in case the first argument does not contradict the argument it is attacking, as in the example **(e)-(f)** showed in Section 3.2.2). Due to the complexity of the annotation, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 80 argument pairs (20 pairs randomly extracted from each of the "complex attacks" data set), and it has the goal to assess the validity of the annotation task (counting when the judges agree on the same annotation). We calculated the inter-annotator agreement for our annotation task in two steps. We *(i)* verify the agreement of the two judges on the argument pairs classification *attacks/unrelated*, and *(ii)* consider only the argument pairs tagged as *attacks* by both annotators, and we verify the agreement between the pairs tagged as *contradiction* and as *null* (i.e. non contradiction). Applying κ to our data, the agreement for the first step is $\kappa = 0.77$, while for the second step $\kappa = 0.71$. Both agreements are satisfactory, although they reflect the higher complexity of the second annotation task (*contradiction/null*), as pointed out in Section 3.2.2.

The distribution of complex attacks in the Debatepedia data set, as resulting after a reconciliation phase carried out by the annotators, is shown in Table 3. As can be noticed, the *mediated* attack is the most frequent type of attack, generating 335 new arguments pairs in the NL sample we considered (i.e., the conditions that allow the application of this kind of complex attacks appear more frequently in real debates). Together with the *secondary* attacks, they appear in the AFs of all the debated topics. On the contrary, *extended* attacks are added in 11 out of 19 topics, and *supported* attacks in 17 out of 19 topics. Considering all the topics, on average only 6 pairs generated from the additional attacks were already present in the original data set, meaning that considering also these attacks is a way to hugely enrich our data set.

Proposed models	**# occ.**	**attacks**		**unrel.**
		+contr(null)	*-contr(null)*	
Supported attacks	47	23	17	7
Secondary attacks	53	29	18	6
Mediated attacks	335	84	148	103
Extended attacks	28	15	10	3

Table 3: Complex attacks distribution in our data.

Figure 3 graphically represents the complex attacks distribution. Considering the first step of the annotation (i.e. *attacks* vs *unrelated*), the figure shows that the latter case is very infrequent, and that (except for the *mediated* attack) on average only 10% of the argument pairs are tagged as *unrelated*. This observation can be considered as a proof of concept of the four theoretical models of complex attacks we analyzed. Due to the

[3] Freely available at `http://bit.ly/VZIs6M`

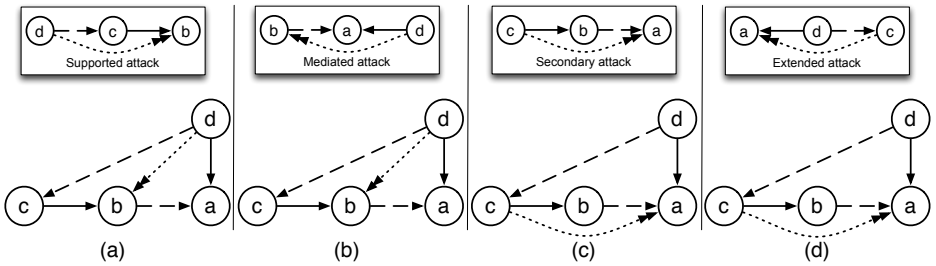

Figure 2: Example of bipolar argumentation framework with the introduction of supported attacks.

fact that the conditions for the application of the *mediated* attacks are verified more often in the data, it has the drawback of generating more unrelated pairs. Still, the number of successful cases is high enough to consider this kind of attack as representative of human interactions. Considering the second step of the annotation (i.e., *attacks* as *contradiction* or *null*), we can see that results are in line with those reported in our first study (Table 2), meaning that also among complex attacks the same distribution is maintained.

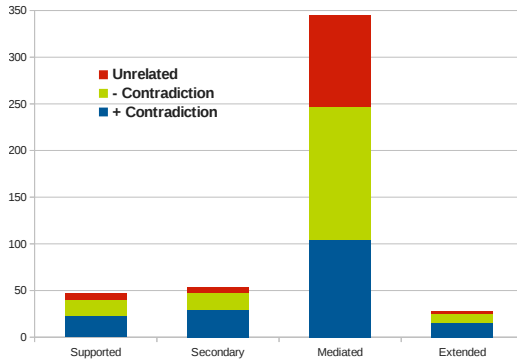

Figure 3: Complex attacks distribution in our data.

4 Concluding remarks

In this paper, we provide a further step towards a better comprehension of the support and attack notions in bipolar argumentation (invoked by the community, e.g. (Cayrol and Lagasquie-Schiex, 2011)) by evaluating them against *naturally occurring data* extracted from NL online debates. The results show that the support relation includes the TE relation, i.e. it is more general (in about 60% of the argument pairs in relation of support, also TE holds). Similarly, the study on the attack-contradiction relations shows that the attack relation is more general than the contradiction (as underlined by (de Marneffe et al., 2008)): in about 70% of the attacks also contradiction holds.

The proposed study shows that the research carried out on semantic inferences in NLP, and argumentation theory in knowledge representation could fruitfully influence each other, raising new open challenges with a significant potential impact on the future interactions among humans and machines. On the one side, NLP provides to the argumentation theory community *i)* textual inference paradigms like TE that make inference algorithms and tools available to automatically process NL arguments, and to detect the semantic relations linking them, and *ii)* annotated natural language corpora that can be investigated in depth to prove the proposed formal models on naturally occurring data. On the other side, argumentation theory can provide to TE, and in general to NLP approaches to semantic inference, a new framework where the semantic relations are not only identified between pairs of textual fragments, but such pairs are also part of an argumentation graph that provides an overall view of the arguments' interactions such that the influences of the arguments on the others emerge, even if they are not direct (see the additional attacks in Section 3.3, and (Berant et al., 2012)'s work on the structural constraints of TE in the context of entailment graphs). Formal models of argumentation are also proposed to check the consistency of a set of information items represented as the nodes of an argumentation graph, allowing for the detection of the precise portions of the graph where the inconsistency arises (e.g., argument a supports and attacks the same argument). This would open new challenges for TE, that in the original definition considers the T-H pairs as "self-contained" (i.e., the meaning of H has to be derived from the meaning of T). On the contrary, in arguments extracted from human linguistic interactions a lot is left implicit (following Grice's conversational Maxim of Quantity), and anaphoric expressions should be solved to correctly assign semantic relations among arguments.

References

Jonathan Berant, Ido Dagan, Meni Adler, and Jacob Goldberger. 2012. Efficient tree-based approximation for entailment graph learning. In *ACL (1)*, pages 117–125.

Guido Boella, Dov M. Gabbay, Leendert van der Torre, and Serena Villata. 2010. Support in abstract argumentation. In *Procs of COMMA, Frontiers in Artificial Intelligence and Applications 216*, pages 111–122.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Procs of ECAI, Frontiers in Artificial Intelligence and Applications 242*, pages 205–210.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artif. Intell.*, 170(11):925–952.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Procs of ECSQARU, LNCS 3571*, pages 378–389.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2010. Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *Int. J. Intell. Syst.*, 25(1):83–109.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2011. Bipolarity in argumentation graphs: Towards a better understanding. In *Procs of SUM, LNCS 6929*, pages 137–148.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(04):i–xvii.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Procs of ACL*.

Phan M. Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358.

Stella Heras, Katie Atkinson, Vicente J. Botti, Floriana Grasso, Vicente Julián, and Peter McBurney. 2010. How argumentation can enhance dialogues in social networks. In *Procs of COMMA, Frontiers in Artificial Intelligence and Applications 216*, pages 267–274.

Thomas K. Landauer, Darrell Laham, Bob Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Procs of CSS*, pages 412–417.

Farid Nouioua and Vincent Risch. 2010. Bipolar argumentation frameworks with specialized supports. In *Procs of ICTAI*, pages 215–218. IEEE Computer Society.

Farid Nouioua and Vincent Risch. 2011. Argumentation frameworks with necessities. In *Procs of SUM, LNCS 6929*, pages 163–176.

Nir Oren and Timothy J. Norman. 2008. Semantics for evidence-based argumentation. In *Procs of COMMA, Frontiers in Artificial Intelligence and Applications 172*, pages 276–284.

Nir Oren, Chris Reed, and Michael Luck. 2010. Moving between argumentation frameworks. In *Procs of COMMA, Frontiers in Artificial Intelligence and Applications 216*, pages 379–390.

Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1:93–124.

Iyad Rahwan and Guillermo Simari, editors. 2009. *Argumentation in Artificial Intelligence*. Springer.

Chris Reed and Floriana Grasso. 2007. Recent advances in computational models of natural argument. *Int. J. Intell. Syst.*, 22(1):1–15.

Adam Wyner and Tom van Engers. 2010. A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *Procs of eGov 2010*.