Argument and Computation, 2013
Vol. 4, No. 3, 209–230, http://dx.doi.org/10.1080/19462166.2013.862303

Taylor & Francis
Taylor & Francis Group

A natural language bipolar argumentation approach to support users in online debate interactions[†]

Elena Cabrio* and Serena Villata

INRIA Sophia Antipolis, Valbonne, France

(*Received 2 August 2013; final version received 31 October 2013*)

With the growing use of the Social Web, an increasing number of applications for exchanging opinions with other people are becoming available online. These applications are widely adopted with the consequence that the number of opinions about the debated issues increases. In order to cut in on a debate, the participants need first to evaluate the opinions of the other users to detect whether they are in favour or against the debated issue. Bipolar argumentation proposes algorithms and semantics to evaluate the set of accepted arguments, given the support and the attack relations among them. Two main problems arise. First, an automated framework to detect the relations among the arguments represented by the natural language (NL) formulation of the users' opinions is needed. Our paper addresses this open issue by proposing and evaluating the use of NL techniques to identify the arguments and their relations. In particular, we adopt the textual entailment (TE) approach, a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. TE is then coupled together with an abstract bipolar argumentation system which allows to identify the arguments that are accepted in the considered online debate. Second, we address the problem of studying and comparing the different proposals put forward for modelling the support relation. The emerging scenario shows that there is not a unique interpretation of the support relation. In particular, different combinations of additional attacks among the arguments involved in a support relation are proposed. We provide an NL account of the notion of support based on online debates, by discussing and evaluating the support relation among arguments with respect to the more specific notion of TE in the NL processing field. Finally, we carry out a comparative evaluation of four proposals of additional attacks on a sample of NL arguments extracted from Debatepedia. The originality of the proposed framework lies in the following point: NL debates are analysed and the relations among the arguments are automatically extracted.

Keywords: argumentation theory; natural language processing; online debates

1. Introduction

In the last years, the Web has changed in the so-called Social Web. The Social Web has seen an increasing number of applications like Twitter,[1] Debatepedia,[2] Facebook[3] and many others, which allow people to express their opinions about different issues. Let us consider, for instance, the following debate published on Debatepedia: the issue of the debate is "Making Internet a right only benefits society". The participants have proposed various pro and con arguments concerning this issue, e.g. a pro argument claims that the Internet delivers freedom of speech, and a con argument claims that the Internet is not as important as real rights like the freedom from slavery. These kinds of debates are composed of tens of arguments in favour or against a proposed issue. The main difficulty for newcomers is to understand the current holding position in the debate, i.e. to understand which are the arguments that are accepted at a certain moment. This difficulty is

*Corresponding author. Email: elena.cabrio@inria.fr

[†]The paper is an extended version of Cabrio and Villata (2012).

twofold: first, the participants have to remember all the different, possibly long, arguments and understand which are the relations among these arguments, and second they have to understand, given these relations, which are the accepted arguments.

In this paper, we answer the following research question: *how to support the participants in natural language (NL) debates to detect which are the relations among the arguments, and which arguments are accepted?* Two kinds of relations connect the arguments in such online debate platforms: a positive relation (i.e. a *support* relation) and a negative relation (i.e. an *attack* relation). To answer to our research question we need to rely on an argumentative framework able to deal with such *bipolar* relations. Dung's (1995) abstract theory defines an argumentation framework as a set of abstract arguments interacting with each other through a so-called *attack* relation. In the last years, several proposals to extend the original abstract theory with a *support* relation have been addressed, leading to the birth of *bipolar argumentation* frameworks (BAFs) (Cayrol & Lagasquie-Schiex, 2005), and the further introduction of a number of *additional attacks* among the arguments (Boella, Gabbay, van der Torre, & Villata, 2010; Cayrol & Lagasquie-Schiex, 2010; Nouioua & Risch, 2011).

Our research question breaks down into the following subquestions:

(1) How to automatically identify the arguments, as well as their relationships, from NL debates?
(2) What is the relation between the notion of support in bipolar argumentation and the notion of textual entailment (TE) in natural language processing (NLP)?

First, we propose to combine NL techniques and Dung-like abstract argumentation to identify and generate the arguments from NL text and then to evaluate this set of arguments to know which are the accepted ones. Starting from the participants' opinions, we detect which ones imply or contradict, even indirectly, the issue of the debate using the TE approach. Beside formal approaches to semantic inference that rely on logical representation of meaning, the notion of TE has been proposed as an applied framework to capture major semantic inference needs across applications in the Computational Linguistics field (Dagan, Dolan, Magnini, & Roth, 2009). The development of the Web has witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. TE is a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. We use TE to automatically identify, from an NL text, the arguments. Second, we adopt bipolar argumentation (Cayrol & Lagasquie-Schiex, 2005) to reason over the set of generated arguments with the aim of deciding which are the accepted ones. Proposals like argumentation schemes (Walton, Reed, & Macagno, 2008), Araucaria (Reed & Rowe, 2004), Carneades (Gordon, Prakken, & Walton, 2007), and ArguMed (Verheij, 1998) use NL arguments, but they ask the participants to indicate the semantic relationship among the arguments, and the linguistic content remains unanalysed. As underlined by Reed and Grasso (2007), "the goal machinery that leads to arguments being automatically generated has been only briefly touched upon, and yet is clearly fundamental to the endeavor". Summarising, we combine the two approaches, i.e. TE and abstract bipolar argumentation, in a framework whose aim is to (i) generate the abstract arguments from the online debates through TE, (ii) build the argumentation framework from the arguments and the relationships returned by the TE module, and (iii) return the set of accepted arguments. We evaluate the feasibility of our combined approach on a data set extracted from a sample of Debatepedia debates.

Second, we study the relation among the notion of support in bipolar argumentation (Cayrol & Lagasquie-Schiex, 2005) and the notion of TE in NLP (Dagan et al., 2009). In the first study of the current work, we assume the TE relation extracted from NL texts as equivalent to a support relation in bipolar argumentation. This is a strong assumption, and in this second part of our work,

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 particular, for addressing this issue we focus both on the relation between support and entailment, and on the relation 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. Moreover, 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 & 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 & Risch, 2010, 2011) (if a supports b and a attacks c, then b attacks c). We investigate the presence and the distribution of these attacks in NL debates 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 originality of the proposed framework consists in the combination of two techniques which need each other to provide a complete reasoning model: TE has the power to automatically identify the arguments in the text and to specify which kind of relation links each couple of arguments, but it cannot assess which are the *winning* arguments. This is addressed by argumentation theory which lacks automatic techniques to extract the arguments from free text. The combination of these two approaches leads to the definition of a powerful tool to reason over online debates. In addition, the benefit of the proposed deeper analysis of the relation among the two notions of support and TE 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 behaviour as close as possible to the human one.

The paper is organised as follows. Section 2 provides an overview on the standard approaches to semantic inference in the NLP field as well as an introduction to TE. Section 3 summarises the basic notions of bipolar argumentation and describes the four kinds of additional attacks we consider in this paper. Section 4 presents our combined framework unifying TE and bipolar argumentation towards the automated detection of the arguments' relations and their acceptability. Section 5 addresses the analysis of the meaning of support and attack in NL dialogues, as well as the comparative study on the existing additional attacks. In Section 6, we compare our framework to the existing related work.

2. NLP approaches to semantic inference

Classical approaches to semantic inference rely on logical representations of meaning that are external to the language itself and are typically independent of the structure of any particular NL. Texts are first translated, or interpreted, into some logical form and then new propositions are inferred from interpreted texts by a logical theorem prover. But, especially after the development of the Web, we have witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. Addressing the inference task by means of logical theorem provers in automated applications aimed at NL understanding has shown several intrinsic limitations (Blackburn, Bos, Kohlhase, & de Nivelle, 2001). As highlighted in Monz and de Rijke (2001), in formal approaches semanticists generally opt for rich (i.e. including at least first-order logic) representation formalisms to capture as many relevant aspects of the meaning as possible, but practicable methods for generating such representations are very rare. The translation of real-world sentences into logic is difficult because of issues such as ambiguity or vagueness (Pinkal, 1995). Moreover, the computational costs of deploying first-order logic theorem prover tools in

real-world situations may be prohibitive, and huge amounts of additional linguistic and background knowledge are required. Formal approaches address forms of deductive reasoning, and therefore often exhibit a too high level of precision and strictness when compared with human judgements, that allow for uncertainties typical of inductive reasoning (Bos & Markert, 2006). While it is possible to model elementary inferences on the precise level allowed by deductive systems, many pragmatic aspects that play a role in everyday inference cannot be accounted for. Inferences that are plausible but not logically stringent cannot be modelled in a straightforward way, but in NLP applications approximate reasoning should be preferred in some cases to having no answers at all.

Especially in data-driven approaches, like the one sought in this work, where patterns are learnt from large-scale naturally occurring data, we can settle for approximate answers provided by efficient and robust systems, even at the price of logic unsoundness or incompleteness. Starting from these considerations, Monz and de Rijke (2001) propose to address the inference task directly at the textual level instead, exploiting currently available NLP techniques. While methods for automated deduction assume that the arguments in input are already expressed in some formal meaning representation (e.g. first-order logic), addressing the inference task at a textual level opens different and new challenges from those encountered in formal deduction. Indeed, more emphasis is put on informal reasoning, lexical semantic knowledge, and variability of linguistic expressions.

The notion of TE has been proposed as an applied framework to capture major semantic inference needs across applications in NLP (Dagan et al., 2009). It is defined as a relation between a coherent textual fragment (the Text T) and a language expression, which is considered as the Hypothesis (H). Entailment holds (i.e. $T \Rightarrow H$) if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. The TE relationship is directional, since the meaning of one expression may usually entail the other, while the opposite is much less certain. Consider the pairs in Examples 2.1 and 2.2.

Example 2.1

T1: *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*
H: *Making Internet a right only benefits society.*

Example 2.2 Continued

T2: *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery, or the right to elementary education.*
H: *Making Internet a right only benefits society.*

A system aimed at recognising TE should detect an inference relation between T1 and H (i.e. the meaning of H can be derived from the meaning of T) in Example 2.1, while it should not detect an entailment between T2 and H in Example 2.2. As introduced before, TE definition is based on (and assumes) common human understanding of language, as well as common background knowledge. However, the entailment relation is said to hold only if the statement in the text licenses the statement in the hypothesis, meaning that the content of T and common knowledge together should entail H, and not background knowledge alone. In this applied framework, inferences are performed directly over lexical–syntactic representations of the texts. Such a definition of TE captures quite broadly the reasoning about language variability needed by different applications aimed at NL understanding and processing, e.g. information extraction (Romano, Kouylekov, Szpektor, Dagan, & Lavelli, 2006) and text summarisation (Barzilay & McKeown, 2005). Differing from the classical

semantic definition of entailment (Chierchia & McConnell-Ginet, 2000), the notion of TE accounts for some degree of uncertainty allowed in applications (see Example 2.1).

In 2005, the PASCAL Network of Excellence started an attempt to promote a generic evaluation framework covering semantic-oriented inferences needed for practical applications, launching the Recognizing Textual Entailment challenge (Dagan et al., 2009), with the aim of setting a unifying benchmark for the development and evaluation of methods that typically address similar problems in different, application-oriented manners. As many of the needs of several NLP applications can be cast in terms of TE, the goal of the evaluation campaign is to promote the development of general entailment recognition engines, designed to provide generic modules across applications. Since 2005, such initiative has been repeated yearly,[4] asking the participants to develop a system that, given two text fragments (the *text* T and the *hypothesis* H), can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other. For pairs where the entailment relation does not hold between T and H, systems are required to make a further distinction between pairs where the entailment does not hold because the content of H is contradicted by the content of T (i.e. *contradiction*, see Example 2.2), and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T (i.e. *unknown*, see Example 2.3). Marneffe, Rafferty, and Manning (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. This three-way judgement task (*entailment* vs. *contradiction* vs. *unknown*) was introduced since RTE-4, while before a two-way decision task (*entailment* vs. *no entailment*) was asked to participating systems. However, the classic two-way task is offered as an alternative also in recent editions of the evaluation campaign (*contradiction* and *unknown* judgements are collapsed into the judgement *no entailment*).

In our work, we consider the three-way scenario to map TE relation with bipolar argumentation, focusing both on the relation between support and entailment, and on the relation between attack and contradiction. As will be discussed in Section 5.2, we consider 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) as *unknown* pairs in the TE framework.

Example 2.3 Continued

T3: *Internet "right" means denying parents' ability to set limits. Do you want to make a world when a mother tells her child: "you cannot stay on the internet anymore" that she has taken a right from him? Compare taking the right for a home or for education with taking the "right" to access the internet.*

H: *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*

The systems submitted to the Recognizing Textual Entailment (RTE) challenge are tested against manually annotated data sets, which include typical examples that correspond to success and failure cases of NLP applications. A number of data-driven approaches applied to semantics have been experimented throughout the years. In general, the approaches still more used by the submitted systems include Machine Learning (typically Support Vector Machines (SVM)), logical inference, cross-pair similarity measures between T and H, and word alignment – for an overview, see Androutsopoulos and Malakasiotis (2010) and Dagan et al. (2009).

3. Bipolar argumentation

This section provides the basic concepts of Dung's (1995) abstract argumentation and bipolar argumentation (Cayrol & Lagasquie-Schiex, 2005).

Definition 3.1 (Abstract argumentation framework, AF) An abstract argumentation framework 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 (1995) presents several acceptability semantics that produce zero, one, or several sets of accepted arguments. Such semantics are grounded on two main concepts called conflict-freeness and defence.

Definition 3.2 (Conflict-free, defence) Let $C \subseteq A$. A set C is *conflict-free* if and only if there exist no $a, b \in C$ such that $a \rightarrow b$. A set C *defends* an argument a if and only if for each argument $b \in A$ if b attacks a then there exists $c \in C$ such that c attacks b.

Definition 3.3 (Acceptability semantics) Let C be a conflict-free set of arguments, and let $\mathcal{D} : 2^A \mapsto 2^A$ be a function such that $\mathcal{D}(C) = \{a | C \text{ defends } a\}$.

- C is *admissible* if and only if $C \subseteq \mathcal{D}(C)$.
- C is a *complete extension* if and only if $C = \mathcal{D}(C)$.
- C is a *grounded extension* if and only if it is the smallest (w.r.t. set inclusion) complete extension.
- C is a *preferred extension* if and only if it is a maximal (w.r.t. set inclusion) complete extension.
- C is a *stable extension* if and only if it is a preferred extension that attacks all arguments in $A \setminus C$.

Roughly, an argument is accepted if all its attackers are rejected, and it is rejected if it has at least an attacker which is accepted.

BAFs, 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 BAF is a labelled 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 3.4 (BAF) A 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. Cayrol and Lagasquie-Schiex (2005, 2010) argue about the emergence of new kinds of attacks from the interaction between attacks and supports in BAF. In the rest of the paper, we will adopt their terminology to refer to additional attacks, i.e. *complex attacks*. In particular, they specify two kinds of complex attacks called *secondary* and *supported* attacks, respectively.

Definition 3.5 (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 $a_1 R_1 \cdots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $\forall i = 1 \cdots n-2, R_i = \dashrightarrow$ and $R_{n-1} = \rightarrow$. A secondary attack for b by a is a sequence $a_1 R_1 \cdots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $R_1 = \rightarrow$ and $\forall i = 2 \cdots n-1, R_i = \dashrightarrow$.

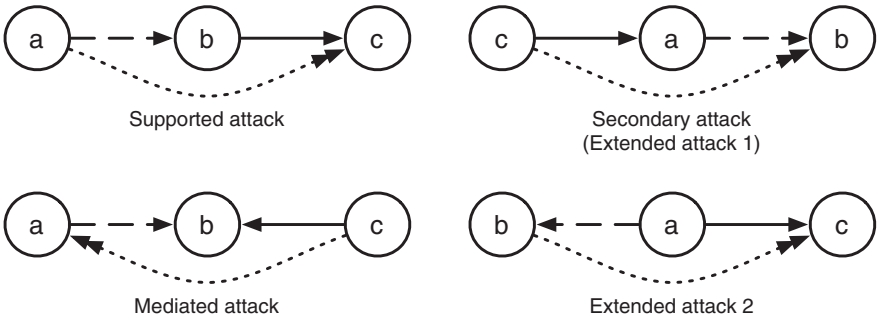

Figure 1. Additional attacks emerging from the interaction of supports and attacks.

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. In this paper, we represent complex attacks using a dotted arrow.

The support relation has been specialised in other approaches where new complex attacks emerging from the combination of existing attacks and supports are proposed. Boella et al. (2010) 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 attacks called *mediated* attacks (Figure 1).

Definition 3.6 (Mediated attacks) Let BAF $= \langle A, \rightarrow, \dashrightarrow \rangle$, where $a, b \in A$. A mediated attack on b by a is a sequence $a_1 R_1 \cdots R_{n-2} a_{n-1}$ 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 $\forall i = 1 \cdots n-2, R_i = \dashrightarrow$.

Nouioua and Risch (2010, 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, 2010), and that the second case is the dual of supported attacks. See Cayrol and Lagasquie-Schiex (2011) for a formal comparison of the different models of support in bipolar argumentation.

Definition 3.7 (Extended attacks) Let BAF $= \langle A, \rightarrow, \dashrightarrow \rangle$, where $a, b \in A$. An extended attack on b by a is a sequence $a_1 R_1 a_2 R_2 \cdots R_n a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $R_1 = \rightarrow$ and $\forall i = 2 \cdots n, R_i = \dashrightarrow$, or a sequence $a_1 R_1 \cdots R_n a_n$ and $a_1 R_p a_p$, $n \geq 2$, with $a_n = a, a_p = b$, such that $R_p = \rightarrow$ and $\forall i = 1 \cdots n, R_i = \dashrightarrow$.

All these models of support in bipolar argumentation address the problem of how to compute the set of extensions from the extended framework providing different kinds of solutions, i.e. introducing the notion of *safety* in BAF (Cayrol & Lagasquie-Schiex, 2005), or computing the extensions in the meta-level (Boella et al., 2010; Cayrol & 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 towards a better understanding of the notion of support and attack in bipolar argumentation.

We are aware that the notion of support is controversial 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 atomic

arguments (Prakken, 2010). However, given the amount of attention bipolar argumentation is receiving in the literature (Rahwan & Simari, 2009), a better account of this kind of frameworks is required.

Another approach to model support has been proposed by Oren and Norman (2008) and Oren, Reed, and Luck (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 do not intend to take a position in this debate. We focus our analysis on the abstract models of bipolar argumentation proposed in the literature (Boella et al., 2010; Cayrol & Lagasquie-Schiex, 2010; Nouioua & Risch, 2011), and we leave as future work the account of support in structured argumentation and the model proposed by Oren and Norman (2008) and Oren et al. (2010).

4. Casting bipolar argumentation as a TE problem

The goal of our work is to propose an approach to support the participants in forums or debates (e.g. Debatepedia, Twitter) to detect which arguments among the ones expressed by the other participants on a certain topic are accepted. As a first step, we need to (i) automatically generate the arguments (i.e. recognise a participant's opinion on a certain topic as an argument), as well as (ii) detect their relation with respect to the other arguments. We cast the described problem as a TE problem, where the T–H pair is a pair of arguments expressed by two different participants in a debate on a certain topic. For instance, given the argument "Making Internet a right only benefits society" (that we consider as H as a starting point), participants can be in favour of it (expressing arguments from which H can be inferred, as in Example 2.1) or can contradict such argument (expressing an opinion against it, as in Example 2.2). Since in debates one participant's argument comes after the other, we can extract such arguments and compare them both w.r.t. the main issue and w.r.t. the other participants' arguments (when the new argument entails or contradicts one of the arguments previously expressed by another participant). For instance, given the same debate as before, a new argument T3 may be expressed by a third participant to contradict T2 (that becomes the new H (H1) in the pair), as shown in Example 4.1.

Example 4.1 Continued

T3: *I have seen the growing awareness within the developing world that computers and connectivity matter and can be useful. It is not that computers matter more than water, food, shelter, and healthcare, but that the network and PCs can be used to ensure that those other things are available. Satellite imagery sent to a local computer can help villages find fresh water, mobile phones can tell farmers the prices at market so they know when to harvest.*

T2 ≡ H1: *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery, or the right to elementary education.*

With respect to the goal of our work, TE provides us with the techniques to identify the arguments in a debate and to detect which kind of relation underlies each couple of arguments. A TE system returns indeed a judgement (entailment or contradiction) on the argument pairs related to a certain topic that are used as input to build the argumentation framework, as described in the

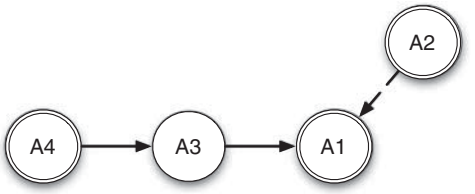

Figure 2. The argumentation framework built from the results of the TE module for Examples 2.1, 2.2, and 4.1.

next section. Example 4.2 presents how we combine TE with bipolar argumentation to compute at the end the set of accepted arguments.

Example 4.2 Continued The TE phase returns the following couples for the NL opinions detailed in Examples 2.1, 2.2, and 4.1:

- T1 entails H
- T2 attacks H
- T3 attacks H1 (i.e. T2)

Given this result, the argumentation module of our framework maps each element to its corresponding argument: H $\equiv A_1$, T1 $\equiv A_2$, T2 $\equiv A_3$, and T3 $\equiv A_4$. The resulting argumentation framework, visualised in Figure 2, shows that the accepted arguments (using admissibility-based semantics) are $\{A_1, A_2, A_4\}$. This means that the issue "Making Internet a right only benefits society" A_1 is considered as accepted. Double bordered arguments are the accepted ones.

4.1. *Experimental setting*

As a case study to experiment the combination of TE and argumentation theory to support the interaction of participants in online debates, we select Debatepedia, an encyclopaedia of pro and con arguments on critical issues. In Section 4.1.1, we describe the creation of the data set of T–H pairs extracted from a sample of Debatepedia topics, in Section 4.1.2 we present the TE system we use, and in Section 4.1.3, we report on obtained results.

4.1.1. *Data set*

To create the data set of argument pairs to evaluate our task, we follow the criteria defined and used by the organisers of RTE (see Section 2). To test the progress of TE systems in a comparable setting, the participants to RTE are provided with data sets composed of T–H pairs involving various levels of entailment reasoning (e.g. lexical, syntactic), and TE systems are required to produce a correct judgement on the given pairs (i.e. to say if the meaning of one text snippet can be inferred from the other). The data available for the RTE challenges are not suitable for our goal, since the pairs are extracted from news and are not linked among each others (i.e. they do not report opinions on a certain topic).

For this reason, we created a data set to evaluate our combined approach focusing on Debatepedia. We manually selected a set of topics (Table 3 column *Topic*) of Debatepedia debates, and for each topic 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 (i.e. when a user's opinion supports or attacks an argument previously expressed by another user, we couple the former with the latter), following the chronological order to maintain the dialogue structure;
(4) the resulting pairs of arguments are then tagged with the appropriate relation, i.e. *attack* or *support*.[5]

Using Debatepedia as case study provides us with already annotated arguments (*pro* ⇒ *entailment*,[6] and *con* ⇒ *contradiction*), and casts our task as a yes/no entailment task. 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). Then, at step 2, we extract all the users' opinions concerning this issue (both pro and con), e.g. (b), (c), and (d):

Example 4.3

(a) *Coca can be classified as a narcotic.*
(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 argument (d) with argument (c) since they follow one another in the discussion (i.e. user expressing argument (c) answers back to user expressing argument (b), so the arguments are concatenated – the same for arguments (d) and (c)).

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)**.

We collected 200 T-H pairs (Table 1), 100 to train and 100 to test the TE system (each data set is composed by 55 entailment and 45 contradiction pairs). The pairs considered for the test set concern completely new topics, never seen by the system (Table 1).

4.1.2. *TE system*

To detect which kind of relation underlies each couple of arguments, we take advantage of the modular architecture of the EDITS system (Edit Distance Textual Entailment Suite) version 3.0, an open-source software package for recognising TE[7] (Kouylekov & Negri, 2010). EDITS implements a distance-based framework which assumes that the probability of an entailment relation

Table 1. The Debatepedia data set used in our experiments.

		#pairs		
Topic	#argum	TOT.	Yes	No
Training set				
Violent games boost aggressiveness	16	15	8	7
China one-child policy	11	10	6	4
Consider coca as a narcotic	15	14	7	7
Child beauty contests	12	11	7	4
Arming Libyan rebels	10	9	4	5
Random alcohol breath tests	8	7	4	3
Osama death photo	11	10	5	5
Privatising social security	11	10	5	5
Internet access as a right	15	14	9	5
Total	109	100	55	45
Test set				
Ground zero mosque	9	8	3	5
Mandatory military service	11	10	3	7
No fly zone over Libya	11	10	6	4
Airport security profiling	9	8	4	4
Solar energy	16	15	11	4
Natural gas vehicles	12	11	5	6
Use of cell phones while driving	11	10	5	5
Marijuana legalisation	17	16	10	6
Gay marriage as a right	7	6	4	2
Vegetarianism	7	6	4	2
Total	110	100	55	45

between a given T–H pair is inversely proportional to the distance between T and H (i.e. the higher the distance, the lower is the probability of entailment).[8] Within this framework the system implements different approaches to distance computation, i.e. both edit distance algorithms (that calculate the T–H distance as the cost of the edit operations, i.e. insertion, deletion, and substitution that are necessary to transform T into H) and similarity algorithms. Each algorithm returns a normalised distance score. At a training stage, distance scores calculated over annotated T–H pairs are used to estimate a threshold that best separates positive from negative examples. Such threshold is then used at a test stage to assign a judgement and a confidence score to each test pair.

4.1.3. *Evaluation*

To evaluate our combined approach, we carry out a two-step evaluation: first, we assess the performances of the TE system to correctly assign the entailment and contradiction relations to the pairs of arguments in the Debatepedia data set. Then, we evaluate how much such performances impact on the application of the argumentation theory module, i.e. how much a wrong assignment of a relation to a pair of arguments is propagated in the argumentation framework.

For the first evaluation, we run EDITS on the Debatepedia training set to learn the model, and we test it on the test set. We tuned EDITS in the following configuration: (i) cosine similarity as the core distance algorithm, (ii) distance calculated on lemmas, and (iii) a stopword list is defined to set no distance between stopwords. We use the system off-the-shelf, applying one of its basic configurations. As future work, we plan to fully exploit EDITS features, integrating background and linguistic knowledge in the form of entailment rules and to calculate the distance between T and H on their syntactic structure.

Table 2. System performances on the Debatepedia data set (precision, recall, and accuracy).

		Train			Test		
	rel	Pr.	Rec.	Acc.	Pr.	Rec.	Acc.
EDITS	Yes	0.71	0.73	**0.69**	0.69	0.72	**0.67**
	No	0.66	0.64		0.64	0.6	
WordOverl.	Yes	0.64	0.65	0.61	0.64	0.67	0.62
	No	0.56	0.55		0.58	0.55	

Note: Best accuracy is reported in bold.

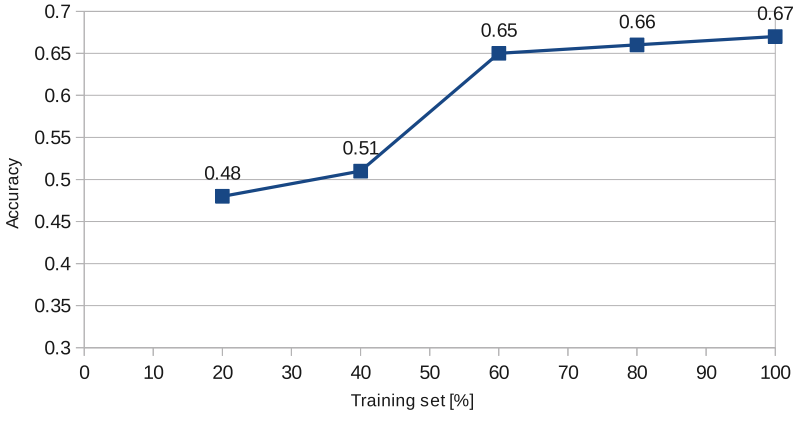

Figure 3. EDITS learning curve on Debatepedia data set.

Table 2 reports on the obtained results both using EDITS and using a baseline that applies a Word Overlap algorithm on tokenised text. Even using a basic configuration of EDITS, and a small data set (100 pairs for training), performances on Debatepedia test set are promising, and in line with performances of TE systems on RTE data sets (usually containing about 1000 pairs for training and 1000 for test). In order to understand if increasing the number of argument pairs in the training set could bring to an improvement in the system performances, the EDITS learning curve is visualised in Figure 3. Note that augmenting the number of training pairs actually improves EDITS accuracy on the test set, meaning that we should consider extending the Debatepedia data set for future work.

As a second step in our evaluation phase, we consider the impact of EDITS performances on the acceptability of the arguments, i.e. how much a wrong assignment of a relation to a pair of arguments affects the acceptability of the arguments in the argumentation framework. We use admissibility-based semantics to identify the accepted arguments both on the correct argumentation framework of each Debatepedia topic (where entailment/contradiction relations are correctly assigned, i.e. the goldstandard) and on the framework generated assigning the relations resulted from the TE system judgements. The precision of the combined approach we propose in the identification of the accepted arguments is on average 0.74 (i.e. arguments accepted by the combined system and by the goldstandard w.r.t. a certain Debatepedia topic) and the recall is 0.76 (i.e. arguments accepted in the goldstandard and retrieved as accepted by the combined system). Its accuracy (i.e. ability of the combined system to accept some arguments and discard some others) is 0.75, meaning that the TE system mistakes in relation assignment propagate in the argumentation framework, but results are still satisfying.

5. Extending the analysis on bipolar argumentation beyond TE

In the previous section, we assumed the TE relation extracted from NL texts as equivalent to the support relation in bipolar argumentation. On closer view, this is a strong assumption. In this second part of our work, we aim at verifying on an extended sample of real data from Debatepedia whether it is always the case that support is equivalent to TE. In particular, for addressing this issue, we focus both on the relations between support and entailment, and on the relations between attack and contradiction. We extend the data set of Section 4.1.1, extracting an additional set of arguments from Debatepedia topics (see Section 5.1). Even if our data set cannot be exhaustive, the methodology we apply for the arguments extraction aims at preserving the original structure of the debate, to make it as representative as possible of daily human interactions in NL.

Two different empirical studies are presented in this section. The first one (Section 5.2) follows the analysis presented in Section 4 and explores the relation among the notion of *support* and *attack* in bipolar argumentation, and the *semantic inferences* as defined in NLP. The second analysis (Section 5.3) starts instead from the comparative study of Cayrol and Lagasquie-Schiex (2011) of the four complex attacks proposed in the literature (see Section 3) and investigates their distribution in NL debates.

5.1. *Data set*

We select the same topics as in Section 4.1.1, since this is the only freely available data set of NL arguments (Table 3, column *Topic*). But since that data set was created respecting the assumption that the TE relation and the support relation are equivalent, in all the previously collected pairs both TE and support relations (or contradiction and attack relations) hold.

In this study we want to move a step further to understand whether it is always the case that support is equivalent to TE (and contradiction to attack). We therefore apply again the extraction methodology described in Section 4.1.1 to extend our data set. In total, our new 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, see Table 3). We consider the obtained data set as representative of human debates in a non-controlled setting (Debatepedia users position their arguments with respect to the others as PRO or CON, the data are not biased), and we use it for our empirical studies.

5.2. *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 & Lagasquie-Schiex, 2011), and the definition of semantic inference in NLP (in particular, the more specific notion of TE) (Dagan et al., 2009).

Basing on the TE definition, an annotator with skills in linguistics has carried out a first phase of annotation of the Debatepedia data set (Section 5.1). 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* judgement 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 in Example 4.3, a correct entailment pair is **(b)** $\Rightarrow$ **(a)**, while a contradiction is **(d)** $\not\Rightarrow$ **(a)**. A *null* judgement is assigned to **(d)–(c)**, since the former argument supports the latter without inferring it. Our data set is an extended version of Cabrio Villata's (2012) one allowing for a deeper investigation.

To assess the validity of the annotation task, we calculate the inter-annotator agreement. Another annotator with skills in linguistics has therefore independently annotated a sample of 100 pairs of the data set. The statistical measure usually used to calculate the inter-rater agreement

Table 3. Debatepedia data set.

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
Privatising 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 legalisation	23	25
Gay marriage as a right	10	10
Vegetarianism	14	13
Total	310	**320**

Note: The total number of collected pairs is reported in bold.

Table 4. Support and TE relations on Debatepedia data set.

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

for categorical items is Cohen's kappa coefficient (Carletta, 1996) which takes into account also agreement occurring by chance. The equation for κ is $\kappa = (\Pr(a) - \Pr(e))/(1 - \Pr(e))$, where $\Pr(a)$ is the relative observed agreement among raters and $\Pr(e)$ is the hypothetical probability of chance agreement. If the raters are in complete agreement then $\kappa = 1$, if there is no agreement among the raters other than what would be expected by chance, $\kappa = 0$. For NLP tasks, the agreement is considered as significant when $\kappa > 0.6$. We calculated the inter-annotator agreement considering the argument pairs tagged as *support* and *attacks* by both annotators, and we verify the agreement between the pairs tagged as *entailment* and as *null* (i.e. no entailment), and as *contradiction* and as *null* (i.e. no contradiction), respectively. Applying κ to our data, the agreement for our task is $\kappa = 0.74$. As a rule of thumb, this is a satisfactory agreement.

Table 4 reports the results of the annotation on our Debatepedia data set, as resulting after a reconciliation phase carried out by the annotators.[9]

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 (e.g. **(d)–(c)** in Example 4.3). 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 Example 5.1.

Example 5.1

(e) *Coca chewing is bad for human health. The decision to ban coca chewing 50 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 relation between support–entailment, the difference between attack and contradiction is more subtle, and it is not always straightforward to say whether an argument attacks another argument without contradicting it. In Example 5.1, we consider that **(e)** does not contradict **(f)** even if it attacks **(f)**, since chewing coca can offer an energy boost, and still be bad for human health. This kind of attack is less frequent than the attacks–contradictions (Table 4).

Considering the three-way scenario 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 also be applied 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) and Heras et al. (2010), argumentation theory should be used as a tool in online discussion applications to identify the relations among the statements and provide a structure to the dialogue to easily evaluate the user's opinions. Starting from the methodology proposed in Section 4 for passing from NL arguments to a BAF, our study demonstrates that applying the TE approach would be productive in the 66% of the Debatepedia data set. Other techniques should then be experimented to cover the other cases, for instance measuring the semantic relatedness of the two propositions using latent semantics analysis techniques (Landauer, Laham, Rehder, & Schreiner, 1997).

5.3. *Second study: complex attacks*

We carry out now a comparative evaluation of the four additional attacks proposed in the literature and investigate their meaning and distribution on the sample of NL arguments.

Basing on the additional attacks (Section 3), and the original AF of each topic in our data set (Table 3), the following procedure is applied: the *supported* (secondary, mediated, and extended, respectively) 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 argument pairs generated from the different types of complex attacks in separate data sets allows us to independently analyse each type and to perform a more accurate evaluation.[10] Figure 4(a)–(d) shows the four AFs resulting from the addition of the complex attacks in the example *Can coca be classified as a narcotic?* Note that the AF in Figure 4(a), where the supported attack is introduced, is the same of Figure 4(b), where the mediated attack is introduced. Notice that, even if the additional attack which is introduced coincide, i.e. d attacks b, this is due indeed 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.

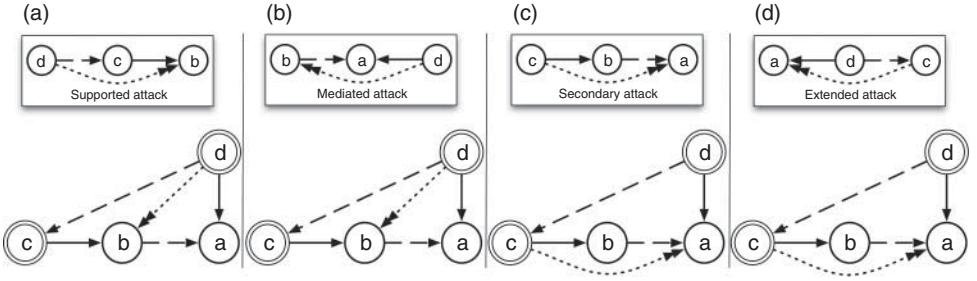

Figure 4. The BAF with the introduction of complex attacks. The top figures show which combination of support and attack generates the new additional attack.

A second annotation phase is then carried out on the data set to verify if the generated argument 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 argument 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 argument pair **(g)–(h)** (Example 5.2) 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.

Example 5.2

(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 Example 5.1). 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 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. no contradiction). Applying κ to our data, the agreement for the first step is $\kappa = 0.77$, while for the second step $\kappa = 0.71$. As a rule of thumb, both agreements are satisfactory, although they reflect the higher complexity of the second annotation (*contradiction*/*null*), as pointed out in Section 5.2.

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 5. As can be noticed, the *mediated* attack is the most frequent type of attack, generating 335 new argument 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 *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 six 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 of NL debates.

Table 5. Complex attack distribution in our data set.

Proposed models	# occ.	Attacks		Unrelated
		+ *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

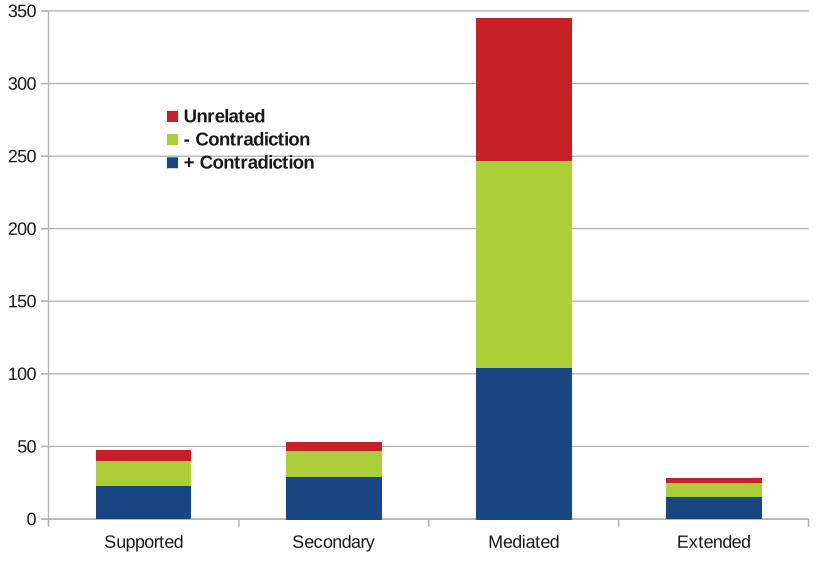

Figure 5. Complex attack distribution in our data set.

Figure 5 graphically represents the complex attack 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 *mediated* attacks) 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 analysed. Due to the 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 4), meaning that also among complex attacks the same distribution is maintained.

6. Related work

Among the set of online debate systems, Debategraph[11] is an online system for debates supporting the incremental development of argument structures, but it is not grounded on argument theory to decide the accepted arguments.

Gilbert (2001) addresses the topic of human/computer argumentation, where the ability to identify and classify various locutions as facts, values, and goals is discussed. The paper does not

present a solution to the problem of automatically generating the arguments from the NL text. The author grounds his observations on the Toulmin (1958) argumentation model.

Chesñevar and Maguitman (2004) use defeasible argumentation to assist the language usage assessment. Their system provides recommendations on language patterns using indices (computed from Web corpora) and defeasible argumentation, where the preference criteria for language usage are formalised as defeasible and strict argumentation rules. The aim of the paper is different from ours. No NL techniques are used to automatically detect and generate the arguments.

Carenini and Moore (2006) present a computational framework for generating evaluative arguments. The framework, based on the user's preferences, produces the arguments following the guidelines of argumentation theory to structure and select evaluative arguments. Then, an NLP step returns the argument in NL. The output of the argumentation strategy is a text plan indicating the propositions to include in the argument and its overall structure. The aim of the paper is different from ours: we do not use NL generation to produce the arguments, but we use TE to detect the arguments in the NL text. We use the word "generation" with the meaning of generation of the abstract arguments from the text, and not with the meaning of NL generation. Concerning argumentation, we use bipolar argumentation to reason over the arguments to identify the accepted ones. We do not address argumentation-based persuasion or planning.

Leite and Martins (2011) envision a self-managing online debating system able to accommodate different kinds of participation of the agents. However, while we re-use Dung's abstract theory, they depart from this approach and defend the view that these debates should provide more than an accepted/rejected classification of the issue at stake. They do not apply NLP techniques to identify the arguments in NL debates.

Wyner and van Engers (2010) present a policy-making support tool based on forums. They propose to couple NLP and argumentation to provide the set of well-structured statements that underlie a policy. Apart from the different goal of this work, there are several points which distinguish our proposal from this one. First, their NLP module guides the participant in writing the input text using Attempt to Controlled English which allows the usage of a restricted grammar and vocabulary. After parsing the text, the sentences are translated to First Order Logic (FOL). We do not have any kind of lexicon or grammar restriction and do not support the participant in writing the text, but automatically extract the arguments from the debates. Second, the inserted statements are associated with a mode indicating the relation between the existing statements and the input statement. Relations among arguments are not manually assigned by participants, but we infer them using TE. Moreover, no evaluation of their framework is provided.

Heras et al. (2010) show how to model the opinions put forward on business-oriented websites using argumentation schemes. The idea is to associate a scheme to each argument to have a formal structure which makes the reasoning explicit. We share the same goal, that is providing a formal structure to online dialogues to evaluate them, but, differently from Heras et al. (2010), in our proposal we achieve this issue using an automatic technique to generate the arguments from NL texts as well as their relations.

Rahwan, Banihashemi, Reed, Walton, and Abdallah (2011) present Avicenna, a Web-based system used to reason about arguments, ranging from automatic argument classification to reason about chained argument structures. In Avicenna, the arguments are inserted by participants through a form, and the participants can decide to attack or support existing arguments, while in our framework participants do not enter arguments: it automatically returns the abstract arguments, the relationships among them highlighting the accepted arguments.

Moens, Boiy, Palau, and Reed (2007) experiment ML approaches to recognise features characterising legal arguments. We adopt a more general framework, i.e. TE (implementable also using ML techniques) to extract open-domain arguments and automatically assign their relations.

Amgoud and Prade (2013) start from a model of argumentation presented in linguistics (Apotheloz, 1993) and try to formalise it using formal argumentation. No automatic treatment of NL arguments is addressed.

7. Conclusions

The research presented in this paper is interdisciplinary. We have integrated in a combined framework an approach from computational linguistics and a technique for non-monotonic reasoning. The aim of this research is to provide the participants of online debates and forums with a framework supporting their interaction with the application. In particular, the proposed framework helps the participants to have an overview of the debates, understanding which are the accepted arguments at time being. The key contribution of our research is to allow the automatic detection and generation of the abstract arguments from NL texts.

First, we adopt a TE approach to inference because of the kind of (noisy) data present on the Web. TE is used to retrieve and identify the arguments, together with the relation between them: the entailment relation (i.e. inference among two arguments) and the attack relation (i.e. contradiction among two arguments). The arguments and their relations are then sent to the argumentation module which introduces the additional attacks. The argumentation module returns the set of acceptable arguments w.r.t. the chosen semantics. We experimented the combined approach on a sample of topics extracted from Debatepedia. We created a data set of 200 pairs of arguments and tested an off-the-shelf open source TE system (i.e. EDITS) on it. The argumentation frameworks built using the relations assigned by the TE system have been evaluated to select the accepted arguments. The accuracy of the combined approach in identifying the arguments in a debate, and to correctly propose to the participant the accepted arguments, is about 75%.

Second, we provide a further step towards a better comprehension of the support and attack notions in bipolar argumentation by evaluating them against *real 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 for the support–entailment, the study on the attack–contradiction relations shows that the attack relation is more general than the contradiction, as underlined by Marneffe et al. (2008): in about 70% of the attacks also contradiction holds. Finally, our study shows that supported attacks do not hold only 14% of the times, secondary attacks do not hold only 11% of the times, mediated attacks do not hold 30% of the times, and extended attacks do not hold only 10% of the times. We point out that the purpose of this paper is not to discuss the criticisms advanced against bipolar argumentation (Amgoud & Prade, 2013), nor to support one model over the other. On the contrary, it is intended as a proof of the concept of the existing models with respect to real data.

Several research lines have to be considered as future research. In particular, the combination of two different techniques will address many open issues in both the research fields of computational linguistics and non-monotonic reasoning. First, the use of NLP to detect the arguments from text will make argumentation theory applicable to reason in real scenarios. We plan to use the TE module to reason on the introduction of the support relation in argumentation theory. Second, given the promising results we obtained, we plan to extend the experimental setting increasing both the pairs of arguments in the Debatepedia data set and to improve the TE system performances to apply our combined approach in other real applications. With this respect, we also plan to carry out an in-depth analysis of the quality of the naturally occurring argumentation in Debatepedia and in other similar online debates platforms as Twitter, or forums, where language complexity is even more challenging for NLP applications. Third, we plan to integrate our combined framework together with a notification module able to send advise to the participants of online debates when a new argument is introduced in the debate against or in favour of her arguments. For a study on the

application of the proposed framework on Wikipedia revisions, see Cabrio, Villata, and Gandon (2013). Finally, we will evaluate if the resulting extensions of BAFs are intuitively correct, i.e. how complex attacks influence the acceptance/rejection of the debated issue, and consider also further relations of structured models (Amgoud & Besnard, 2009). Related to this issue, we plan to address a user evaluation (following the idea of Rahwan, Madakkatel, Bonnefon, Awan, & Abdallah, 2010) to check with the debate participants whether or not they agree with the evaluation returned by the BAF, i.e. whether they agree with the set of *winning* arguments identified by argumentation semantics.

Notes

1. http://twitter.com/.
2. http://idebate.org/.
3. http://www.facebook.com/.
4. http://aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment.
5. The data set is freely available at http://bit.ly/debatepedia_ds.
6. Here we consider only arguments implying another argument. Arguments "supporting" another argument, but not inferring it will be discussed in Section 5.
7. http://edits.fbk.eu/.
8. In previous RTE challenges, EDITS always ranked among the 5 best participating systems out of an average of 25 systems and is one of the few RTE systems available as open source http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool.
9. In this phase, the annotators discuss the results to find an agreement on the annotation to be released.
10. Data sets freely available for research purposes at http://bit.ly/VZIs6M.
11. http://debategraph.org.

References

Amgoud, L., & Besnard, P. (2009, September). Bridging the gap between abstract argumentation systems and logic. In *Proceedings of the 3rd international conference on Scalable Uncertainty Management (SUM), LNCS 5785* (pp. 12–27). Washington, DC: Springer Verlag.

Amgoud, L., & Prade, H. (2013). Can AI models capture natural language argumentation? *International Journal of Cognitive Informatics and Natural Intelligence*, *6*, 19–32.

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*, 135–187.

Apotheloz, D. (1993). The function of negation in argumentation. *Journal of Pragmatics*, *19*(1), 23–38.

Barzilay, R., & McKeown, K. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, *31*(3), 297–327.

Blackburn, P., Bos, J., Kohlhase, M., & de Nivelle., H. (2001). Inference and computational semantics. *Studies in Linguistics and Philosophy, Computing Meaning*, *77*(2), 11–28.

Boella, G., Gabbay, D.M., van der Torre, L.W.N., & Villata, S. (2010, September). Support in abstract argumentation. In *Proceedings of Computational Models of Argument (COMMA), frontiers in artificial intelligence and applications 216* (pp. 111–122). Desenzano del Garda, Italy: IOS Press.

Bos, J., & Markert, K. (2006, October). When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the 2nd PASCAL workshop on recognizing textual entailment*. Venice.

Cabrio, E., & Villata, S. (2012, August). Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI), frontiers in artificial intelligence and applications 242* (pp. 205–210). Montpellier: IOS Press.

Cabrio, E., Villata, S., & Gandon, F. (2013, May). A support framework for argumentative discussions management in the web. In *Proceedings of the 10th international conference on the semantic web: Semantics and big data (ESWC), LNCS 7882* (pp. 412–426). Montpellier: Springer Verlag.

Carenini, G., & Moore, J.D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence*, *170*, 925–952.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*, 249–254.

Cayrol, C., & Lagasquie-Schiex, M.C. (2005, July). On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), LNCS 3571* (pp. 378–389). Barcelona: Springer Verlag.

Cayrol, C., & Lagasquie-Schiex, M.C. (2010). Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *International Journal of Intelligent Systems*, *25*, 83–109.

Cayrol, C., & Lagasquie-Schiex, M. (2011, October). Bipolarity in argumentation graphs: Towards a better understanding. In *Proceedings of the 5th international conference on Scalable Uncertainty Management (SUM), LNCS 6929* (pp. 137–148). Dayton, OH: Springer Verlag.

Chesñevar, C.I., & Maguitman, A. (2004, August). An argumentative approach to assessing natural language usage based on the web corpus. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)* (pp. 581–585). Valencia: IOS Press.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics* (2nd ed.). Cambridge, MA: MIT Press.

Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering Journal*, *15*, i–xvii.

Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*, 321–358.

Gilbert, M. (2001, May). Getting good value. Facts, values, and goals in computational linguistics. In *Proceedings of the International Conference on Computational Science (ICCS), LNCS 2073* (pp. 989–998). San Francisco, CA: Springer Verlag.

Gordon, T.F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*, 875–896.

Heras, S., Atkinson, K., Botti, V.J., Grasso, F., Julián, V., & McBurney, P. (2010, September). How argumentation can enhance dialogues in social networks. In *Proceedings of computational models of argument (COMMA), frontiers in artificial intelligence and applications 216* (pp. 267–274). Desenzano del Garda, Italy: IOS Press.

Kouylekov, M., & Negri, M. (2010, July). An open-source package for recognizing textual entailment. In *Proceedings of the association for computational linguistics, system demonstrations (ACL)* (pp. 42–47). Uppsala: The Association for Computer Linguistics.

Landauer, T.K., Laham, D., Rehder, B., & Schreiner, M.E. (1997, December). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of CSS* (pp. 412–417). San Diego, CA.

Leite, J., & Martins, J. (2011, July). Social abstract argumentation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2287–2292). Barcelona: IJCAI/AAAI.

Marneffe, M.D., Rafferty, A., & Manning, C. (2008, June). Finding contradictions in text. *Proceedings of the 46th annual meeting of the Association for Computational Linguistics (ACL)*, Columbus, OH.

Moens, M., Boiy, E., Palau, R.M., & Reed, C. (2007, June). Automatic detection of arguments in legal texts. *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)* (pp. 225–230). Stanford, California, USA.

Monz, C., & de Rijke, M. (2001, June). Light-weight entailment checking for computational semantics. In *Proceedings Inference in Computational Semantics (ICoS-3)* (pp. 59–72). Siena.

Nouioua, F., & Risch, V. (2010, October). Bipolar argumentation frameworks with specialized supports. In *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 215–218). Arras, France: IEEE Computer Society.

Nouioua, F., & Risch, V. (2011, October). Argumentation frameworks with necessities. In *Proceedings of the 5th international conference Scalable Uncertainty Management (SUM), LNCS 6929* (pp. 163–176). Dayton, OH: Springer Verlag.

Oren, N., & Norman, T.J. (2008, May). Semantics for evidence-based argumentation. In *Proceedings of the international conference on computational models of argument (COMMA), frontiers in artificial intelligence and applications 172* (pp. 276–284). Toulouse: IOS Press.

Oren, N., Reed, C., & Luck, M. (2010, September). Moving between argumentation frameworks. In *Proceedings of the international conference on computational models of argument (COMMA), frontiers in artificial intelligence and applications 216* (pp. 379–390). Desenzano del Garda, Italy: IOS Press.

Pinkal, M. (1995). *Logic and lexicon: The semantics of the indefinite*. Studies in Linguistics and Philosophy, 56. Springer Verlag.

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

Rahwan, I., Banihashemi, B., Reed, C., Walton, D., & Abdallah, S. (2011). Representing and classifying arguments on the semantic web. *Knowledge Engineering Review*, *26*, 487–511.

Rahwan, I., Madakkatel, M.I., Bonnefon, J.F., Awan, R.N., & Abdallah, S. (2010). Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, *34*, 1483–1502.

Rahwan, I., & Simari, G. (Eds.). (2009). *Argumentation in artificial intelligence*. Springer.

Reed, C., & Grasso, F. (2007). Recent advances in computational models of natural argument. *International Journal of Intelligent Systems*, *22*, 1–15.

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, *13*, 961–980.

Romano, L., Kouylekov, M.O., Szpektor, I., Dagan, I., & Lavelli, A. (2006, April). Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of the 11st conference of the European chapter of the association for computational linguistics (EACL)* (pp. 409–416). Avignon: The Association for Computer Linguistics.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Verheij, B. (1998, December). Argumed – a template-based argument mediation system for lawyers and legal knowledge based systems. In *Proceedings of the 11th international conference on legal knowledge and information systems (JURIX)* (pp. 113–130). Amsterdam.

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

Wyner, A., & van Engers, T. (2010, September). A framework for enriched, controlled on-line discussion forums for e-government policy-making. *Proceedings of the annual international e-government conference (eGov)*, Bilbao, Spain.