Abstract Dialectical Frameworks for Text Exploration

Elena Cabrio[1] and Serena Villata[2]
[1] *University of Nice Sophia Antipolis, Sophia Antipolis, France*
[2] *CNRS, I3S Laboratory, Sophia Antipolis, France*

Keywords: Argumentation, Textual Entailment, Natural Language Processing.

Abstract: Textual Entailment (TE) systems aim at recognizing the relations of *entailment* or *non entailment* holding between two text fragments (i.e. a pair). The identified TE pairs are considered as independent one from the others. However, in the latest years TE systems have been challenged against a number of real world application scenarios like analyzing costumers interactions about a service, or analyzing online debates. These applications have underlined the need to move from TE pairs to TE graphs where pairs are no more independent. Moving from single pairs to graphs has the advantage of providing an overall view of the topic discussed in the text. The challenge here is to define ways to exploit such graph-based representation for text exploration. In the literature, some approaches apply abstract argumentation theory to compute the accepted arguments of a debate, but they present a number of drawbacks, e.g., the *non entailment* relation and the *attack* relation in abstract argumentation are assumed to be equivalent, but this is not always the case. In this paper, we define bipolar entailment graphs, i.e., graphs whose nodes are text fragments and the edges represent the *entailment* or *non entailment* relations. We adopt abstract dialectical frameworks to define acceptance conditions for the nodes such that the resulting framework returns us *relevant* information for our text exploration task. Experimental evaluation shows the feasibility of our approach.

1 INTRODUCTION

In the last ten years, the Textual Entailment (TE) framework (Dagan et al., 2009) has gained popularity in Natural Language Processing (NLP) applications like information extraction and question answering, providing a suitable model for capturing major semantic inference needs at textual level, taking into account the language variability. Given a pair of textual fragments, a TE system assigns an *entailment* or a *non entailment* relation to the pair. However, in real world scenarios as analyzing costumers' interactions about a service or a product, or online debates, these pairs extracted from the interactions cannot be considered as independent. This means that they need to be collected together into a single graph, e.g., all the reviews about a certain service are collected together to understand which are the overall problems/merits of the service.[1] This combination of TE pairs into a unique graph aims at supporting text exploration, whose goal is the extraction of specific information from users interactions evaluated as relevant in a particular domain or task. The challenge is thus to propose an automated framework able to compute such relevant information starting from the TE pairs returned by the system and collected into a graph.

In this paper, we answer the research question:

- How to guide text exploration by highlighting relevant information?

Differently from standard entailment graphs (Berant et al., 2010; Mehdad et al., 2013) where the nodes are connected by entailment relations only, in this paper we consider *bipolar entailment graphs* (BEG), where the nodes are the text fragments of TE pairs, and both relations returned by TE systems (i.e., *entailment* and *non entailment*) are considered as the graph links. A recent proposal by (Cabrio and Villata, 2012) suggests that TE pairs can be collected together to construct an abstract argumentation framework (Cayrol and Lagasquie-Schiex, 2013; Dung, 1995) where the *entailment* relation is mapped with the *support* relation in argumentation, and the *non entailment* relation is mapped with the *attack* relation. Argumentation theory (Dung, 1995) is used to compute the set of accepted arguments in the online debates they analyze. While we believe that strong connections hold

[1] As discussed also in the keynote talk of the Joint Symposium on Semantic Processing (http://jssp2013.fbk.eu/)

Cabrio, E. and Villata, S.
Abstract Dialectical Frameworks for Text Exploration.
DOI: 10.5220/0005699100850095
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 85-95
ISBN: 978-989-758-172-4

between TE and argumentation theory, we detect the following drawbacks in their combined approach: *i)* the *non entailment* relation is considered as equivalent to a contradiction and directly translated into an attack relation. This is not always the case: *non entailment* means that the two text spans are either *unrelated* or contradicting each other; *ii)* the support relation affects arguments' acceptability only if supported arguments are also attacked (new attacks are introduced when a support holds (Cayrol and Lagasquie-Schiex, 2013)), making the resulting framework more complex; and *iii)* applying standard acceptability semantics (Dung, 1995) to TE graphs does not give the possibility to express detailed task-dependent conditions to be satisfied, in order to have the arguments accepted.

Our research question breaks down into the following sub-questions:

- How to cast bipolar entailment graphs in the argumentation setting such that the semantics of the relations is maintained?
- How to define specific arguments' acceptance conditions such that information we consider as relevant in our task is extracted?

First, we answer the research questions by adopting *abstract dialectical frameworks* (ADF) (Brewka and Woltran, 2010; Brewka et al., 2013), a generalization of Dung's abstract argumentation frameworks where different kinds of links among statements are represented. We cast bipolar entailment graphs in abstract dialectical frameworks where the links represent entailment and non entailment.

Second, considering positive (entailing) and negative (non entailing) links, and the weights assigned to such links by the TE system, we define and evaluate two acceptance conditions which allow us to extract in an automated way the set of arguments, i.e., text fragments, relevant for our text exploration task.

The goal of the proposed framework is to highlight the information that is relevant to explore (i.e. to understand, and in a certain sense, to summarize) humans interactions in natural language (e.g. in a debate, or in a reviewing service). Our proposal is a natural language based knowledge representation framework grounded on natural language constructs rather than on a formal pre-defined terminology. On the one side we provide an automated way to compute relevant information, and on the other side we apply abstract dialectical frameworks to a real application where texts are the primary source of knowledge.

In the remainder of the paper, Section 2 compares the proposed approach to the related work. Section 3 presents the TE framework. Section 4 introduces ADFs and the two acceptance conditions we define. Experimental setting is described in Section 5. Conclusions end the paper.

2 RELATED WORK

The term *entailment graph* is not new in the literature, and it has been firstly introduced by Berant et al. (Berant et al., 2010) as a structure to model entailment relations between propositional templates. The nodes of an entailment graph are propositional templates, i.e., a path in a dependency tree between two arguments of a common predicate (Lin and Pantel, 2001). In a dependency parse, such a path passes through the predicate; a variable must appear in at least one of the argument positions, and each sense of a polysemous predicate corresponds to a separate template (and a separate graph node): $X \xleftarrow{subj} \textit{treat\#1} \xrightarrow{obj} Y \textit{ and } X \textit{ subj} \xleftarrow{subj} \textit{treat\#1} \xrightarrow{obj} \textit{nausea}$ are propositional templates for the first sense of the predicate *treat*. An edge (u,v) represents the fact that template u entails template v. (Berant et al., 2010) assume a user interested in retrieving information about a target concept (e.g., *nausea*). The proposed approach automatically extracts from a corpus the set of propositions where *nausea* is an argument, and learns an entailment graph over propositional templates derived from the extracted propositions.

While (Berant et al., 2010; Berant et al., 2012) model the problem of learning entailment relations between predicates represented as propositional templates as a graph learning problem (to search for the best graph under a global transitivity constraint), we collect both entailment and non entailment relations returned by the system to use both of them during the computation of relevant information. In the context of the topic labeling task, (Mehdad et al., 2013) propose to build a multidirectional entailment graph over the phrases extracted for a given set of sentences (covering the same topic). Since many of such phrases include redundant information which are semantically equivalent but vary in lexical choices, they exploit the entailment graphs to discover if the information in one phrase is semantically equivalent, novel, or more/less informative with respect to the content of the other phrase.

Also the combination of argumentation theory and NLP is not new, and some existing works combine NLP and argumentation theory (Chesñevar and Maguitman, 2004; Carenini and Moore, 2006; Moens et al., 2007; Wyner and van Engers, 2010; Feng and Hirst, 2011; Amgoud and Prade, 2012) with different purposes, ranging from policy making support up to recommendations on language patterns using indices, to automated arguments generation. However, only few of them (Carenini and Moore, 2006; Moens et al., 2007; Feng and Hirst, 2011) actually process the textual content of the arguments, but their goals, i.e., arguments generation (Carenini and Moore, 2006), and arguments classification in texts (Moens et al., 2007; Feng and Hirst, 2011) differ from ours.

Moreover, systems like Avicenna (Rahwan et al., 2011), Carneades (Gordon et al., 2007), Araucaria (Reed and Rowe, 2004) (based on argumentation schemes (Walton et al., 2008)), and ArguMed (Verheij, 1998) use natural language arguments, but the text remains unanalyzed as users are requested to indicate the kind of relationship holding between two arguments. Finally, approaches like (Leite and Martins, 2011; Gabbriellini and Torroni, 2013b; Heras et al., 2013) show the added value of applying argumentation theory to understand online discussions and user opinions in decision support and business oriented websites. Again texts here are not the source of knowledge, and the linguistic content is not analyzed. All these approaches show the need to make the two communities communicate and jointly address such kind of open issues.

Up to our knowledge, the only work which tries to combine TE with argumentation theory is (Cabrio and Villata, 2012). The drawbacks of this work have been previously detailed. For sake of completeness, we have to mention that they (Cabrio and Villata, 2012) are aware about the first drawback we identified in their approach, i.e., the fact that the non entailment relation is mapped to the attack relation even if the meaning of the two is different, and they present a data-driven comparison of the meanings of entailment/support and non entailment/attack in (Cabrio and Villata, 2013). However, the drawback still holds, and a more general framework is required to obtain a proper combination of TE and argumentation.

The added value of using argumentation theory in on-line discussions and user reviews to support decision making on business oriented websites has been shown by (Gabbriellini and Santini, 2015), while an interesting approach to support argumentative discussions on social networks, and more precisely on Twitter, has been explored by (Gabbriellini and Torroni, 2012; Gabbriellini and Torroni, 2013a). We share with these approaches the adoption of argumentation theory to support intelligent interactions with other users or big amount of data.

Finally, in the last years, the argument mining research topic has become more and more relevant in the Artificial Intelligence and Natural Language Processing communities, as witnessed by the success of the 'Argument Mining' workshop[2]. An interesting approach that is worth mentioning in particular has been recently presented by (Lippi and Torroni, 2015). The authors propose a method that exploits structured parsing information to detect claims without resorting to contextual information. Even if the goal of the two approaches is different, they go in the same direction of developing supporting systems for users who interact with big amount of data and need to be guided to achieve an intelligent exploration experience.

3 BIPOLAR ENTAILMENT GRAPHS

This section introduces the Textual Entailment framework (Section 3.1), and its extension into bipolar entailment graphs (Section 3.2).

3.1 Textual Entailment

In the NLP field, the notion of Textual Entailment (Dagan et al., 2009) refers to a directional relation between two textual fragments, termed *Text (T)* and *Hypothesis (H)*, respectively. The relation holds (i.e. $T \Rightarrow H$) whenever the truth of one text fragment follows from another text, as interpreted by a typical language user. The TE relation is directional, since the meaning of one expression may usually entail the other, while entailment in the other direction is much less certain. Consider the pairs in Examples 1, 2, and 3:

Example 1.
T (id=3)*: People should be at liberty to treat their bodies how they want to. Indeed, people are allowed to eat and drink to their detriment and even death, so why shouldn't they be able to harm themselves with marijuana use? This is, of course, assuming that their use does not harm anyone else.*
H (id=1)*: Individuals should be free to use marijuana. If individuals want to harm themselves, they should be free to do so.*

Example 2 (Continued)**.**
T (id=2)*: Even if marijuana's effects were isolated to the individual, there is room for the state to protect individuals from harming themselves.*
H (id=1)*: Individuals should be free to use marijuana. If individuals want to harm themselves, they should be free to do so.*

[2] https://www.cs.cornell.edu/home/cardie/naacl-2nd-arg-mining/

Example 3 (Continued).
T (id=4): *Individuals should be at liberty to experience the punishment of a poor choice.*

H (id=2): *Even if marijuana's effects were isolated to the individual, there is room for the state to protect individuals from harming themselves.*

In Example 1, we can identify an *entailment* relation between T and H (i.e. the meaning of H can be derived from the meaning of T), in Example 2, T *contradicts* H, while in Example 3, even if the topic is the same, the truth of H cannot be verified on the bases of the information present in T (i.e. the relation is said to be *unknown*).[3] The notion of TE has been proposed as an applied framework to capture major semantic inference needs across applications in NLP (e.g. information extraction, text summarization, and reading comprehension systems) (Dagan et al., 2009). The task of recognizing TE is therefore carried out by automatic systems, mainly implemented using Machine Learning techniques (typically SVM), logical inference, cross-pair similarity measures between T and H, and word alignment.[4] While entailment in its logical definition pertains to the meaning of language expressions, the TE model does not represent meanings explicitly, avoiding any semantic interpretation into a meaning representation level. Instead, in this applied model inferences are performed directly over lexical-syntactic representations of the texts. TE allows to overcome the main limitations showed by formal approaches (where the inference task is carried out by logical theorem provers), i.e. *(i)* the computational costs of dealing with huge amounts of available but noisy data present in the Web; *(ii)* the fact that formal approaches address forms of deductive reasoning, exhibiting a too high level of precision and strictness as compared to human judgments, that allow for uncertainties typical of inductive reasoning. But while methods for automated deduction assume that the arguments in input are already expressed in some formal 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.

3.2 From Pairs to Graphs

As defined in the previous section, TE is a directional relation between two textual fragments. However, in various real world scenarios, these pairs cannot be considered as independent. This means that they need to be collected together into a single graph. A new framework involving *entailment graphs* is therefore needed, where the semantic relations are not only identified between pairs of textual fragments, but such pairs are also part of a graph that provides an overall view of the statements' interactions, such that the influences of some statements on the others emerge. Therefore, we introduce the notion of *bipolar entailment graphs* (*BEG*), where two kinds of edges are considered, i.e., entailment and non entailment, and nodes are the text fragments of TE pairs.

Definition 1 (Bipolar Entailment Graph). *A bipolar entailment graph is a tuple* $BEG = \langle T, E, NE \rangle$ *where*

- T *is a set of text fragments;*
- $E \subseteq T \times T$ *is an entailment relation between text fragments;*
- $NE \subseteq T \times T$ *is a non entailment relation between text fragments.*

This opens 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.

4 TEXT EXPLORATION THROUGH ARGUMENTATION

In this section, we first introduce abstract dialectical frameworks (Section 4.1), and then we describe which acceptability measures we choose for our text exploration task (Section 4.2).

4.1 Abstract Dialectical Frameworks

Abstract dialectical frameworks (Brewka and Woltran, 2010) have been introduced as a generalization of Dung-style abstract argumentation frameworks (Dung, 1995) where each node is associated with an acceptance condition. The slogan of abstract dialectical frameworks is: *ADF = dependency graphs + acceptance conditions*, meaning that,

[3] In the two-way classification task, contradiction and unknown relations are collapsed into a unique relation, i.e. *non entailment*.

[4] (Dagan et al., 2009) provides an overview of the recent advances in TE.

in contrast with Dung frameworks where links between nodes represent the type of relationship called *attack*, in this framework different dependencies can be represented in a flexible way.

An ADF is a directed graph whose nodes represent statements which can be accepted or not. The links between the nodes represent dependencies: the status (i.e., accepted, not accepted) of a node s depends only on the status of its parents $par(s)$, i.e., those nodes connected to s by a direct link. Each node s is then associated to an *acceptance condition* C_s which specifies the exact conditions under which argument s is accepted. C_s is a function assigning to each subset of $par(s)$ one of the values *in* or *out*, where *in* means that these arguments are accepted and *out* means that they are rejected. Roughly, if for $R \subseteq par(s)$ we have $C_s(R) = in$, this means that s will be accepted if the nodes in R are accepted and those in $par(s) \setminus R$ are rejected.

Definition 2 (Abstract Dialectical Framework (Brewka and Woltran, 2010))**.** *An abstract dialectical framework is a tuple* $D = \langle S, L, C \rangle$ *where*

- *S is a set of statements (i.e., nodes);*
- *$L \subseteq S \times S$ is a set of links;*
- *$C = \{C_s\}_{s \in S}$ is a set of total functions $C_s : 2^{par(s)} \rightarrow \{in, out\}$, one for each statement s. C_s is called the acceptance condition of s.*

For instance, Dung-style argumentation frameworks are associated to the ADF $D_{Dung} = \langle Args, att, C \rangle$ where the acceptance conditions for all nodes $s \in S$ is $C_s(R) = in$ if and only if $R = \emptyset$, and $C_s(R) = out$ otherwise. An example of an abstract dialectical framework from (Brewka and Woltran, 2010) is visualized in Figure 1, where grey nodes are the accepted arguments, and acceptance conditions are expressed as propositional formulas over the nodes. For more details see (Brewka and Woltran, 2010).

(Brewka and Woltran, 2010) underline that ADF acceptance conditions can be defined also through *positive* and *negative* weights associated to links. In particular, they introduce weighted ADFs presenting their usefulness in the specific context of legal argumentation, i.e., modeling five standards of proof. In this paper, we start from weighted ADFs presented in (Brewka and Woltran, 2010), and we adapt them to represent our bipolar entailment graphs. Note that weighted argumentation frameworks have been studied also by (Dunne et al., 2011), where weights are used for handling inconsistencies, but there weights are not exploited to compute the acceptance or rejection of the arguments. The advantage of using ADFs to model bipolar entailment graphs, in contrast with the approach proposed in (Cabrio and Villata, 2012), is that the resulting "bipolar" argumentation graphs are not forced to interpret the negative weighted links as being attacks and therefore leading to a misconception about the meaning of the non entailment relation in TE.

4.2 Extracting Meaningful Information using ADF

To explore texts searching for information which satisfies specific constraints and shows certain features, we adopt weighted abstract dialectical frameworks (Brewka and Woltran, 2010), and we define two acceptance conditions such that they allow us to select, starting from a bipolar entailment graph, only the information we are looking for. First, we define a general weighted ADF (to which we map *BEG*s) where an additional function is introduced to associate each link to a weight, similarly to what was proposed in (Brewka and Woltran, 2010).

Definition 3 (Weighted Abstract Dialectical Frameworks)**.** *A weighted abstract dialectical framework is a tuple* $D = \langle S, L, C, v \rangle$ *where*

- *S is a set of nodes;*
- *$L \subseteq S \times S$ is a set of links;*
- *$C = \{C_s\}_{s \in S}$ is a set of total functions $C_s : 2^{par(s)} \rightarrow \{in, out\}$, one for each statement s. C_s is called the acceptance condition of s;*
- *$v : L \rightarrow W$ is a function associating weights to the links, where W is a set of weights.*

Mapping a BEG into a weighted ADF, we can highlight two kinds of possible weights in bipolar entailment graphs: *i)* qualitative weights, where we distinguish between *positive* vs. *negative* weights $W = \{+, -\}$, i.e., we consider the entailment links as associated to a positive weight and non entailment links as associated to a negative weight, and *ii)* numerical weights, where we exploit the weights the TE system assigns to each link as its confidence, i.e., we consider a range $W \in [-1, 1]$ such that the more the link weight approaches -1, the more the system is confident it is a non entailment relation and the more the link weight approaches 1, the more the system is confident it is an entailment relation. Figure 1 shows an example of a weighted ADF, where C_s is described.

Starting from the defined weighted ADFs, we have now to define the acceptance conditions we want to adopt to guide the selection of the nodes in the graph that we consider as relevant in our task. We consider two use cases for text exploration: *(a)* a huge online debate composed by several arguments, and we

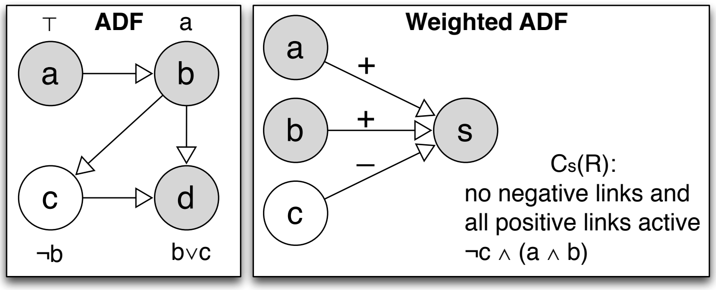

Figure 1: Examples of ADF and weighted ADF together with the acceptance conditions defined for nodes.

want to retrieve the arguments that are entailed by at least one accepted statement and no negative link is directed against them from accepted statements; and *(b)* a set of users' interactions about a service have to be explored in order to retrieve those statements which are highly entailed by other statements in the BEG, and not much non entailed by other statements (i.e., if the difference of their weights is above a certain threshold). These two domain independent acceptance conditions represent our *heuristics* to retrieve inside huge bipolar entailment graphs, the set of information satisfying the goal of our text exploration task.

The two acceptance conditions are formalized as follows:

1. $C_s(R) = in$ if and only if

$$\exists r \in R : v((r,s)) \in \{+\} \wedge \forall t \in R : v((t,s)) \notin \{-\} \quad (1)$$

2. $C_s(R) = in$ if and only if, given $r,t \in R$,

$$\max v_+((r,s)) - |\max v_-((t,s))| > k \quad (2)$$

where k is a certain threshold.

The first acceptance condition models use case *(a)*: statement s is accepted if and only if R contains no node with a negative link towards s and at least one node with a positive link towards s, i.e., no node not entailing s and at least one node entailing s. The second acceptance condition models use case *(b)*: statement s is accepted if and only if the difference between the maximal positive weight and the absolute value of the maximal negative weight is above a given threshold k. Concerning those nodes which have no incident links (i.e., $par(s) = \emptyset$), we apply the following acceptance condition: C_s is in (constant function). Note that we do not claim that these are the only possible acceptance conditions for identifying relevant information during text exploration in BEGs. We define such acceptance conditions because they provide us with the information satisfying our text exploration features. However, weighted ADFs applied to text exploration based on bipolar entailment graphs provide a flexible framework such that more complex acceptance conditions can be defined depending on the kind of information to be retrieved.

5 EXPERIMENTAL SETTING

This section evaluates the automated framework we propose to support text exploration. As a first step, we run a TE system to assign the entailment and the non entailment relations to the pairs of arguments. Then, a bipolar entailment graph is built, where the arguments are the nodes of the graph, and the automatically assigned relations correspond to the links of the graphs. Finally, we adopt the abstract dialectical frameworks to define acceptance conditions for the nodes of the bipolar entailment graph. The dataset of argument pairs on which we run the experiments is described in Section 5.1, while the framework evaluation is reported in Section 5.2.

5.1 Dataset

We experiment our framework on the Debatepedia dataset[5] (described in (Cabrio and Villata, 2012)). It is composed of 200 pairs, balanced between entailment and non entailment pairs, and split into a training set (100 pairs), and a test set (100 pairs). The pairs are extracted from a sample of Debatepedia[6] debates, an encyclopedia of pro and con arguments on critical issues (e.g. China one-child policy, vegetarianism, gay marriages). To the best of our knowledge, it is the only available dataset of T-H pairs that can be represented as bipolar entailment graphs.

Since (Cabrio and Villata, 2012) show on a learning curve that augmenting the number of training pairs actually improves the TE system performances

[5] The Recognizing Textual Entailment (RTE) data are not suitable for our goal, since the pairs are not interconnected (i.e. they cannot be transformed into argumentation graphs)

[6] http://idebate.org/

on the test set, we decided to contribute to the extension of the Debatepedia data set manually annotating 60 more pairs (30 entailment and 30 non entailment pairs). We followed the methodology described in (Cabrio and Villata, 2012) for the annotation phase, and we added the newly created pairs to the original training set. We consider this enriched dataset of 260 pairs as the goldstandard in our experiments (where entailment/non entailment relations are correctly assigned), against which we will compare the TE system performances.

Starting from the pairs in the Debatepedia dataset, we then build a bipolar entailment graph for each of the topic in the dataset (12 topics in the training set and 10 topics in the test set, listed in (Cabrio and Villata, 2012)). The arguments are the nodes of the graph, and the relations among the arguments correspond to the links of the graphs.

To create the goldstandards to check the validity of the two proposed acceptance conditions, we separately applied both conditions on the bipolar entailment graphs built using manually annotated relations. In particular, for the second acceptance condition that consider the weights assigned on the links (see Section 4), we consider the max weight of 1 to be attributed to the entailment link (maximal confidence on the entailment relation assignment), and the max weight of -1 to be attributed to the non entailment link (maximal confidence on the non entailment relation assignment).

We are aware that the dataset we used is smaller than the datasets provided in RTE challenges[7], but we consider it as a representative test set to prove the validity of our approach.

5.2 Evaluation

We carry out a two-step evaluation of our framework: first, we assess the TE system accuracy in correctly assigning the entailment and the non entailment relations to the pairs of arguments in the dataset. Then, we evaluate how much such accuracy impacts on ADF graphs, i.e. how much a wrong assignment of a relation to a pair of arguments is propagated in the ADF by the acceptance conditions.

To detect which kind of relation underlies each couple of arguments, we experiment the EXCITEMENT Open Platform (EOP)[8], that provides a generic architecture for a multilingual textual inference platform. We tested the three state-of-the-art entailment algorithms in the EOP (i.e., BIUTEE (Stern and Dagan, 2012), TIE and EDITS (Kouylekov and Negri, 2010)) on Debatepedia dataset, experimenting several different configurations, and adding knowledge resources.

The best results for the first evaluation step on Debatepedia are obtained with BIUTEE, adopting the configuration that exploits all available knowledge resources (e.g. WordNet, Wikipedia, FrameNet) (see Table 1). BIUTEE follows the transformation-based paradigm, which recognizes TE by converting the text into the hypothesis via a sequence of transformations. Such sequence is referred to as a *proof*, and is performed over the syntactic representation of the text (i.e. the text parse tree). A transformation modifies a given parse tree, resulting in a generation of a new parse tree, which can be further modified by subsequent transformations. The main type of transformations is the application of entailment-rules (Bar-Haim et al., 2007) (e.g. lexical rules, active/passive rules, coreference).

As baseline in this first experiment we use a token-based version of the Levenshtein distance algorithm, i.e. EditDistanceEDA in the EOP, as shown in Table 1. In this table, we do not report the results of the TIE system as it is not relevant with respect to the present evaluation, as we fixed EditDistanceEDA as our baseline and the best performing system for our task in the EOP is BIUTEE. The obtained results are in line with the average systems performances at RTE ($\sim$0.65 F-measure[9]).

As a second step of our evaluation, we consider the impact of the best TE configuration 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 ADF. We use the acceptance conditions we defined in Section 4 to identify the accepted arguments both on *i)* the goldstandard entailment graphs of Debatepedia topics (described in Section 5.1), and *ii)* on the graphs generated using the relations and the weights assigned by BIUTEE on Debatepedia (since it is the system that obtained the best performances, see Table 1).

BIUTEE allows many types of transformations, by which an hypothesis can be proven from any text. Given a T-H pair, the system finds a proof which generates H from T, and estimates the proof validity (Stern and Dagan, 2012). Finding such a proof is a sequential process, conducted by a search algorithm. In each step of the proof construction the system examines all the possible transformations that can be applied, generates new trees by applying the selected transformations, and calculates their costs by con-

[7] http://bit.ly/RTE-challenge

[8] http://hltfbk.github.io/Excitement-Open-Platform/

[9] The F-measure is a measure of accuracy. It considers both the precision and the recall of the test to compute the score.

Table 1: First step evaluation (results on Debatepedia test set, i.e. 100 pairs). Systems are trained on Debatepedia training set (160 pairs).

EOP configuration	**Accuracy**	**Recall**	**Precision**	**F-measure**
BIUTEE	0.71	0.94	0.66	0.78
EditDistanceEDA	0.58	0.61	0.59	0.59

structing appropriate feature-vectors for them. Eventually, the search algorithm finds the (approximately) lowest cost proof. If the proof cost is below a threshold (automatically learned on the training set, for details see (Stern and Dagan, 2011)), then the system concludes that T entails H. The inverse of this cost is normalized as a score between 0 (where T and H are completely different) and 1 (where T and H are identical), and returned as output. In other words, the score returned by the system indicates how likely it is that the obtained proof is valid, i.e., the transformations along the proof preserve entailment from the meaning of T.

In order to apply the second acceptance condition described in Section 4 using the scores returned by BIUTEE as the weights on the links between nodes, we need to have positive values (from 0 to 1) corresponding to the confidence of BIUTEE in assigning the entailment relation to the pair, and negative values (from 0 to -1) corresponding to the confidence of BIUTEE in assigning a non entailment relation to the pair. Since the scores that BIUTEE returns are normalized between 0 and 1, where the threshold learned on the Debatepedia training set is set to 0.5, we need to shift such scores on the scale demanded by such acceptance condition, setting the threshold to 0 and normalizing the scores produced by BIUTEE accordingly. In this new scale, *i)* the more the system is confident that there is a non entailment relation between two arguments, the more its score (i.e. the link weight) approaches -1; *ii)* the more the system is confident that there is an entailment relation, the more its score (i.e. the link weight) approaches 1; *iii)* the more the system is uncertain about the assigned relation, the more the system score (i.e. the link weight) approaches 0 (both on the negative and on the positive scale).

Table 2 reports on the results of this second evaluation phase, where we evaluate the impact of BIUTEE on the arguments acceptability, adopting admissible based semantics, with respect to a goldstandard where the relations on the links have been assigned by human annotators (Section 5.1). In general, the TE system mistakes in relation assignment propagate in the argumentation framework, but results are still satisfying.

We are aware that in Debatepedia entailment graphs the error propagation is also limited by *i)* their size (see Table 2, column *avg # links per graph*); and *ii)* the heuristic we applied in computing the arguments acceptability, according to which the arguments that have no negative incident links are accepted, augmenting the number of the accepted nodes in the graphs. Concerning time complexity, the weighted ADF module takes ∼1 second to analyze a weighted ADF of 100 pairs, returning the relevant arguments with respect to the selected acceptance condition.[10] The results reported in Table 2 cannot be strictly compared with the results shown in (Cabrio and Villata, 2012), since the underlying role of the entailment relation in the selection of the accepted argument is different. In this paper, we do not address a comparison with the existing ADF software, such as DIAMOND and QADF[11], as the purpose of the present paper is not to evaluate the performances in computing ADFs, but the goodness of our system in retrieving natural language arguments for topics exploration. However, we plan as future research to adopt such systems for computing the acceptability of the arguments, and to evaluate their performances with respect to our specific task. Note that this evaluation is not intended to evaluate the performances of argumentation systems to compute the acceptability of the arguments[12], but it is meant to show the accuracy of the combined system (i.e., TE plus ADFs) in detecting the arguments satisfying the specified features, so that it can be exploited for a text exploration task.

In general, we consider the results we obtained experimenting our framework on the Debatepedia dataset as promising, fostering further research in this direction. An analysis of arguments returned by the acceptability conditions has been addressed, and results show that the selected arguments contain relevant information for the topics exploration.

In Figure 2, ADF_1 shows the weighted ADF resulting from the BEG whose text fragments are presented in Section 3, together with the nodes selected through the first acceptance condition. Note that statement "Individuals should be free to use marijuana. If in-

[10] Complexity results for ADFs have been studied by (Brewka and Woltran, 2010).

[11] http://www.dbai.tuwien.ac.at/research/project/adf/

[12] We refer the interested reader to the results of the First International Competition on Computational Models of Argumentation (Thimm and Villata, 2015).

Table 2: Results of the second evaluation (Debatepedia test set). Precision (avg): arguments accepted by the automatic system and by the goldstandard with respect to an entailment graph; recall (avg): arguments accepted in the goldstandard and retrieved as accepted by the automatic system.

Acceptance condition	**# graphs**	**avg # links per graph**	**Precision**	**Recall**	**F-measure**
First	10	9.1	0.89	0.98	0.93
Second	10	9.1	0.894	0.98	0.95

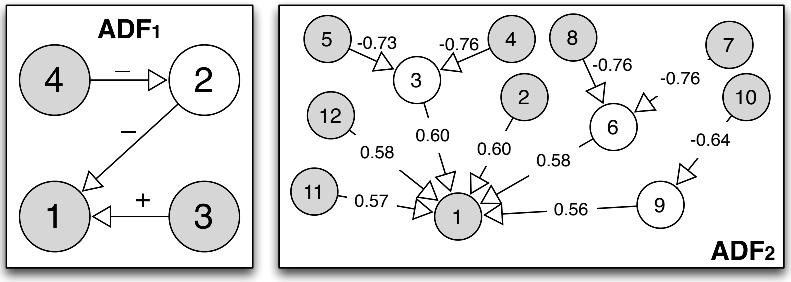

Figure 2: Two examples from our dataset (ADF_1 - positive/negative weights, ADF_2 - numerical weights).

dividuals want to harm themselves, they should be free to do so" is selected as it has an incident negative link but coming from a rejected argument, and it is entailed by "People should be at liberty to treat their bodies how they want to. Indeed, people are allowed to eat and drink to their detriment and even death, so why shouldn't they be able to harm themselves with marijuana use? [...]". In Figure 2, ADF_2 shows the weighted ADF we obtain for the whole ADF about the topic "Gas Vehicles" from our dataset, where the links are weighted with the confidence the TE system associates to the assigned relations. In this case, we first assign to the arguments the acceptability degree computed following the formula of the second acceptance condition, and if the computed value is above the threshold the argument is selected, i.e., it is evaluated as *in*, otherwise it is discarded.

6 CONCLUSIONS

We have introduced the notion of *bipolar entailment graph* where the pairs identified by the classical TE framework are collected together into a single graph. The advantage of moving from pairs to a graph lies in the fact that the graph provides a structured view of the text supporting text exploration tasks. In particular, we propose to exploit abstract dialectical frameworks to perform such tasks: we define acceptance conditions for the nodes such that the framework returns us relevant information for our text exploration task. Relying on ADFs ensures to our framework high flexibility in defining the kind of nodes we look for, i.e., the acceptance conditions, and allows us to overcome some drawbacks highlighted in similar approaches in the literature (Cabrio and Villata, 2012). Experiments on the Debatepedia dataset using state of the art TE systems to automatically assign the inference relations between the statements are promising, fostering further research in this direction. Both the enriched Debatepedia dataset (260 pairs), and the generated ADF are available for research purposes.[13]

As for future work, we will test our framework on a dataset built of customer interactions, where further acceptance conditions may become necessary to retrieve other information in the texts. Moreover, we will study how to modify the acceptance condition to consider the fact that the relations assigned to a pair by the system with a low confidence (around 0) are more uncertain than those assigned with a higher confidence. More specifically, we will consider to associate to the confidence values (from -1 to 1) a probability distribution, to improve the system ability in assigning the semantic relation to the pair, depending on the presence of the entailment relation. An in depth user evaluation of the arguments returned after applying the acceptance conditions for the text exploration task is an ongoing work. Finally, we plan to explore the adoption of GRAPPA (Brewka and Woltran, 2014), a semantical framework that allows to define Dung-style semantics for arbitrary labelled graphs, proposing acceptance functions based on multisets of labels. This framework could allow to simplify the definition of the acceptance functions thanks to the introduced pattern language, enhancing the automated evaluation of our framework.

[13] http://bit.ly/DebatepediaExtended

REFERENCES

Amgoud, L. and Prade, H. (2012). Can AI models capture natural language argumentation? *IJCINI*, 6(3):19–32.

Bar-Haim, R., Dagan, I., Greental, I., and Shnarch, E. (2007). Semantic inference at the lexical-syntactic level. In *AAAI*, pages 871–876.

Berant, J., Dagan, I., Adler, M., and Goldberger, J. (2012). Efficient tree-based approximation for entailment graph learning. In *ACL (1)*, pages 117–125.

Berant, J., Dagan, I., and Goldberger, J. (2010). Global learning of focused entailment graphs. In *ACL*, pages 1220–1229.

Brewka, G., Strass, H., Ellmauthaler, S., Wallner, J. P., and Woltran, S. (2013). Abstract dialectical frameworks revisited. In *IJCAI*.

Brewka, G. and Woltran, S. (2010). Abstract dialectical frameworks. In *KR*.

Brewka, G. and Woltran, S. (2014). GRAPPA: A semantical framework for graph-based argument processing. In Schaub, T., Friedrich, G., and O'Sullivan, B., editors, *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 153–158. IOS Press.

Cabrio, E. and Villata, S. (2012). Natural language arguments: A combined approach. In *ECAI*, pages 205–210.

Cabrio, E. and Villata, S. (2013). A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230.

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

Cayrol, C. and Lagasquie-Schiex, M.-C. (2013). Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, 54(7):876–899.

Chesñevar, C. I. and Maguitman, A. (2004). An argumentative approach to assessing natural language usage based on the web corpus. In *ECAI*, pages 581–585.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(Special Issue 04):i–xvii.

Dung, P. (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.

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., and Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artif. Intell.*, 175(2):457–486.

Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In *ACL*, pages 987–996.

Gabbriellini, S. and Santini, F. (2015). A micro study on the evolution of arguments in amazon.com's reviews. In Chen, Q., Torroni, P., Villata, S., Hsu, J. Y., and Omicini, A., editors, *PRIMA 2015: Principles and Practice of Multi-Agent Systems - 18th International Conference, Bertinoro, Italy, October 26-30, 2015, Proceedings*, volume 9387 of *Lecture Notes in Computer Science*, pages 284–300. Springer.

Gabbriellini, S. and Torroni, P. (2012). Large scale agreements via microdebates. In Ossowski, S., Toni, F., and Vouros, G. A., editors, *Proceedings of the First International Conference on Agreement Technologies, AT 2012, Dubrovnik, Croatia, October 15-16, 2012*, volume 918 of *CEUR Workshop Proceedings*, pages 366–377. CEUR-WS.org.

Gabbriellini, S. and Torroni, P. (2013a). Arguments in social networks. In Gini, M. L., Shehory, O., Ito, T., and Jonker, C. M., editors, *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13, Saint Paul, MN, USA, May 6-10, 2013*, pages 1119–1120. IFAAMAS.

Gabbriellini, S. and Torroni, P. (2013b). Netarg: an agent-based social simulator with argumentative agents. In *AAMAS*, pages 1365–1366.

Gordon, T., Prakken, H., and Walton, D. (2007). The carneades model of argument and burden of proof. *Artif. Intell.*, 171(10-15):875–896.

Heras, S., Atkinson, K., Botti, V. J., Grasso, F., Julián, V., and McBurney, P. (2013). Research opportunities for argumentation in social networks. *Artif. Intell. Rev.*, 39(1):39–62.

Kouylekov, M. and Negri, M. (2010). An open-source package for recognizing textual entailment. In *ACL (System Demonstrations)*, pages 42–47.

Leite, J. and Martins, J. (2011). Social abstract argumentation. In *IJCAI*, pages 2287–2292.

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Lippi, M. and Torroni, P. (2015). Context-independent claim detection for argument mining. In Yang, Q. and Wooldridge, M., editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191. AAAI Press.

Mehdad, Y., Carenini, G., Ng, R. T., and Joty, S. R. (2013). Towards topic labeling with phrase entailment and aggregation. In *HLT-NAACL*, pages 179–189.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *ICAIL*, pages 225–230.

Rahwan, I., Banihashemi, B., Reed, C., Walton, D., and Abdallah, S. (2011). Representing and classifying arguments on the semantic web. *Knowledge Eng. Review*, 26(4):487–511.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(4):961–980.

Stern, A. and Dagan, I. (2011). A confidence model for syntactically-motivated entailment proofs. In *RANLP*, pages 455–462.

Stern, A. and Dagan, I. (2012). Biutee: A modular open-source system for recognizing textual entailment. In *ACL (Demo)*, pages 73–78.

Thimm, M. and Villata, S. (2015). System descriptions of the first international competition on computational models of argumentation (iccma'15). *CoRR*, abs/1510.05373.

Verheij, B. (1998). Argumed - a template-based argument mediation system for lawyers and legal knowledge based systems. In *JURIX*, pages 113–130.

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

Wyner, A. and van Engers, T. (2010). A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *eGov*.