Generating Abstract Arguments: a Natural Language Approach

Elena CABRIO and Serena VILLATA
INRIA Sophia Antipolis, France

Abstract. Many argumentation tools have been proposed nowadays to support the users in on-line social discussions. However, the main drawback of these tools is that they do not cope with the automatic generation of the arguments from the natural language discussions of the users. In this paper, we propose to use a technique from computational linguistics, namely textual entailment, to generate in an automatic way the abstract arguments from the dialogues. The abstract arguments as well as their relationships are then structured in an argumentation graph to evaluate the dialogue as a whole. The success criteria of the proposed approach is that it is able to represent the dynamics of the dialogues among users allowing to find the use of argumentation natural enough to be really adopted.

Keywords. Textual entailment, abstract argumentation theory, bipolar argumentation

1. Introduction

Argumentation theory allows people to formalize and structure their discussions and dialogues. The latest years have seen an increasing number of applications supporting online discussions. Some of them, like Debategraph[1] or Debatepedia[2] propose a kind of structure to guide the interaction of the users. For instance, Debategraph supports the incremental development of argument structure using a graph visualization, and Debatepedia supports the insertion of pros and cons arguments with respect to a central issue. These systems allow the users to "use" argumentation theory in a rather natural way, but they have the drawback of not being grounded on argumentation theory to elicit the accepted arguments. On the one hand, the argumentation community has started to propose new ways to embed argumentation theory for enhancing social networks dialogues [9], social bookmarking and idea-linking [13] or forum discussions [15]. The issue is that, in these works, the arguments the users put forward are constrained in some way, e.g., by associating a scheme to each argument [11,9], or by constraining arguments insertion with a restricted grammar and vocabulary [15]. On the other hand, proposals like Araucaria [12], Carneades [8], and ArguMed [14] use natural language arguments, but they ask the user to indicate the semantic relationship among the arguments. The main drawback of all these approaches is that the linguistic content remains unanalyzed, leading to an unnatural use of argumentation thus discouraging its adoption. In this paper, we agree

[1] `http://debategraph.org`
[2] `http://www.debatepedia.org/`

with the observation from Heras et al. [9] that on-line discussions applications should provide tools to identify attack and support statements and provide a structure to the dialogue in such a way that the user's opinions can be easily evaluated. Here, we answer the following research question: *how to automatically identify attack and support statements from natural language dialogues, to evaluate and structure the whole dialogue?*

To answer the research question, we propose to use natural language techniques to generate the abstract arguments as well as their relationships from natural language text. The result of this generation is then sent to a Dung-like abstract argumentation [7] module which has the aim to provide a formal structure and evaluation to the dialogue as a whole. Starting from the users' opinions, we detect which ones imply or contradict, even indirectly, the issue of the dialogue using the textual entailment approach (TE), an applied framework to capture major semantic inference needs across applications in the Computational Linguistics field [6]. We use TE to automatically identify, from a natural language text, the abstract arguments and their relationships, i.e., attack or support. Finally, we propose an experimental evaluation of the proposed framework which witnesses the feasibility of the approach using a dataset built from Debatepedia.

There are other approaches which couple natural language processing and argumentation. Chasnevar and Maguitman [5] use defeasible argumentation to assist the language usage assessment. Their system provides recommendations on language patterns and defeasible argumentation. They do not use natural language techniques to automatically detect and generate the arguments. Carenini and Moore [3] present a complete 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. The aim of this paper is different from the aim of ours: we do not use natural language generation to produce the arguments, but we use TE to detect the arguments in natural language text. We use the word "generation" with the meaning of generation of the abstract arguments from the text, and not with the meaning of natural language generation. We use the computational abstract model proposed by Dung to reason over the arguments to identify the accepted ones. Wyner and van Engers [15] present a policy making support tool based on forums. They propose to couple natural language processing 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 user in writing the text using a restricted grammar and vocabulary. We do not have any kind of lexicon or grammar restriction. Second, the inserted statements are associated with a mode which indicates the relationship between the existing statements and the input statement. We do not ask the user to explicit the relationship among the arguments, we infer them using TE. Moreover, no evaluation of their framework is provided. Heras et al. [9] show how to model the opinions put forward on business oriented websites using argumentation schemes. We share the same goal, that is providing a formal structure to on-line dialogues to evaluate them, but, differently from [9], in our proposal we achieve this issue using an automatic technique to generate the arguments from natural language texts as well as their relationships.

The reminder of the paper is as follows. Section 2 presents the textual entailment module and explains the proposed approach using an example on the debate "Solar energy is economically sound" from Debatepedia. In Section 3, we describe the experimental setting as well as its evaluation.

2. The textual entailment approach to semantic inference

In the Natural Language Processing (NLP) field, the notion of textual entailment 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 Example 1 and 2 extracted from Debatepedia.

Example 1
T1: Solar energy is abundant. Every minute, enough energy arrives at planet Earth to meet human energy demands for a year. It is, therefore, the most abundant energy source available to humans. This abundance makes it an economic gem.
H: Solar energy is economically sound.

Example 2 (Continued)
T2: Compared to fossil fuels, sunlight is a weak energy source because the radiation strength is "diluted" by the time the rays reach earth. This makes its collection more difficult and expensive. In general, more high technology, equipment, and land-area are required with solar energy to produce the same amount of energy as other resources. This makes it more challenging and expensive.
H: Solar energy is economically sound.

In Example 1, we can identify an inference relation between T1 and H (i.e. the meaning of H can be derived from the meaning of T1), while in Example 2, T2 contradicts H. The notion of TE has been proposed [6] as an applied framework to capture major semantic inference needs across applications in NLP (e.g. information extraction, text summarization, and reading comprehension systems). The task of recognizing TE is therefore carried out by automatic systems, mainly implemented basing on Machine Learning techniques (typically SVM), logical inference, cross-pair similarity measures between T and H, and word alignment.[3] 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. The TE framework allows to overcome the main limitations showed by formal approaches (where the inference task is carried out by means of 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 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. Natural language inference systems exploit therefore the achievements reached in NLP tasks such as syn-

[3] *Dagan et al. (2009)* [6] provides an overview of the recent advances in TE.

tactic parsing, computational lexical semantics and coreference resolution, in order to tackle the more challenging problems of sentence-level semantics.

In this paper, we propose an approach to detect the arguments as well as their relationships to discover which are the accepted ones. As a first step, we need to *(i)* generate the arguments (i.e. automatically recognize a user opinion on a certain topic as an argument), as well as *(ii)* detect their relationships with respect to the other arguments. We therefore cast the described problem as a TE problem, where the T-H pair is a pair of arguments expressed by two different users in a dialogue on a certain topic. For instance, given the argument "Solar energy is economically sound" (that we consider as H as a starting point), users can be in favor of it (expressing arguments from which H can be inferred, as in Example 1), or can contradict such argument (expressing an opinion against it, as in Example 2). Since in dialogues, one user's argument comes after the other, we can extract such arguments and compare them both with respect to the main issue, and with respect to the other users' arguments (when the new argument entails or contradicts one of the arguments previously expressed by another user). For instance, given the same debate as before, a new argument T3 may be expressed with the goal of contradicting T1 (that becomes the new H (called H1) in the pair), as shown in Example 3.

Example 3 (Continued)
T3: Solar panels cannot produce energy at night like other alternatives. Coal-electricity and hydroelectricity can both operate 24/7. Solar power, however, can only operate during the day-time. This means that solar power's energy yield is smaller relative to the capital investment.
T1 ≡ H1: Solar energy is abundant. Every minute, enough energy arrives at planet Earth to meet human energy demands for a year. It is, therefore, the most abundant energy source available to humans. This abundance makes it an economic gem.

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

The textual entailment step returns a set of couples of the kind: argument A_i is in contradiction with argument A_j, or argument A_i entails argument A_j. The aim of the argumentation module is to provide the user with a structured view of the whole dialogue, i.e., of the different opinions, and to show which are the accepted ones, w.r.t. a particular Dung's semantics. We consider two relations among arguments. First, we map the contradiction with the attack relation in abstract argumentation. Second, the entailment relation is viewed as a support relation among abstract arguments. The introduction of the support relation in abstract argumentation is a controversial issue. Many proposals have been addressed with the aim of extending abstract argumentation frameworks with the support relation, i.e., bipolar argumentation frameworks $BAF = \langle \mathscr{A}, \rightarrow, \Rightarrow \rangle$, where $\Rightarrow$ is a binary support relation over $\mathscr{A}$ [4]. In this paper, we do not intend to take a position in this dispute. We choose to follow the model of support proposed by Boella et al. [1], where the introduction of a support between two arguments may lead to the insertion of additional attacks among the arguments. In particular, given that $a \Rightarrow b$, then if there is argument c attacking argument b, we introduce an additional attack, called *mediated* attack, $c \rightarrow b$. Mediated attacks [1] are a kind of attacks which holds in the natural language examples we analyzed, due to the fact that the support we detect using TE is

based on inference. Let us consider now the argumentation framework which structures the dialogue detailed in Example 1, 2, and 3.

Example 4 (Continued) *The textual entailment module returns the following couples for the natural language opinions:*

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

Given this result, the argumentation module maps each element to its corresponding argument: $H \equiv A_1$, $T1 \equiv A_6$, $T2 \equiv A_3$, and $T3 \equiv A_7$. The resulting BAF shows that the accepted arguments among these four arguments (using admissibility-based semantics) are $\{A_1, A_2, A_4\}$. Plain arrows represent attacks, double arrows represent entails, and black arguments are the accepted ones. This means that the issue "Solar energy is economically sound" A_1 is accepted. Figure 1 visualizes the BAF of the entire debate about the subject "Solar energy is economically sound" on Debatepedia, as it is returned by the TE module. For clarity of the picture, we visualize only a subset of mediated attacks.

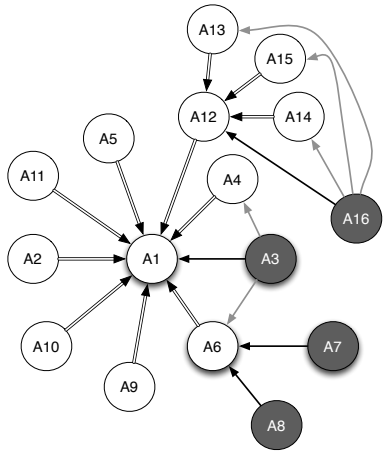

Figure 1. The *BAF* built from TE results for the entire debate "Solar energy is economically sound". Grey attacks are (some of) the mediated attacks used to compute the accepted arguments. Black arguments are the accepted ones.

3. Experimental Setting

To create the data set of arguments' pairs to evaluate our task, we follow the criteria defined and used by the organizers of the Recognizing Textual Entailment Challenges (RTE).[4] To test the progress of TE systems in a comparable setting, the participants to the RTE Campaign are provided with annotated data sets composed of T-H pairs involving various levels of entailment reasoning (e.g. lexical, syntactic), and TE systems are re-

[4] `http://www.nist.gov/tac/2010/RTE/`

quired to produce a correct judgment on the given pairs (i.e. to understand if the meaning of one text snippet can be inferred from the other). The data sets available for the RTE challenges are not suitable for our goal, since the pairs are extracted from newspapers and are not linked among each other, i.e., they do not report opinions on a certain topic.

As a case study to experiment the use of TE to generate the arguments from online dialogues, we select Debatepedia and we created a data set[5] to evaluate our framework as described in [2]. We manually selected a set of topics of Debatepedia debates, and for each topic we coupled all the pros and cons arguments both with the "main" argument, i.e., the title of the debate, as in Example 1 and 2, and/or with other arguments, e.g., Example 3. Using Debatepedia provides us with already annotated arguments (*pro* $\Rightarrow$ *entailment*[6], and *cons* $\Rightarrow$ *contradiction*), and casts our task as a yes/no entailment task. For the experiments described in this paper, we improved our data set [2] with more training pairs, and we made it balanced with respect to the percentage of yes/no judgments. In total, we collected 300 T-H pairs, 200 used to train the TE system (100 entailment, 100 contradiction pairs), and 100 to test it (50 entailment and 50 contradiction pairs). We consider these annotated data sets as the *goldstandard*, i.e. the reference data set (pre-defined by the evaluators) to which the performances of our combined system are compared. The pairs collected for the test set concern completely new topics, never seen by the system, and are provided in their unlabeled form as input.

To detect which kind of relation underlies each couple of arguments, we take advantage of the modular architecture of the EDITS[7] system (Edit Distance Textual Entailment Suite) version 3.0, an open-source software package for recognizing TE [10]. EDITS implements a distance-based framework which assumes that the probability of an entailment relation 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, providing 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 normalized distance score (a number between 0 and 1). 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. The threshold, which is stored in a model, is used at a test stage to assign an entailment judgment and a confidence score to each test pair (i.e. if the distance calculated over a certain test pair is below the threshold, the pair is judged as entailment, while if the distance is above the threshold the pair is judged as contradiction).

We carry out a two-step evaluation: first, we assess the performances of the TE system, i.e. EDITS, to correctly assign the entailment and the contradiction relations to the pairs of arguments in the data set. Then, we evaluate how much such performances impact on the *BAF* used to structure the data, i.e. how much a wrong assignment of a relation to a pair of arguments is propagated in the *BAF*.

---

[5] The data set is freely available at `http://bit.ly/debatepedia_ds`

[6] We considered only favorable arguments that imply another argument, leaving for future work arguments "supporting" another argument, but that do not infer it.

[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 two RTE systems available as open source `http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`

We run EDITS on the Debatepedia training set to learn the model, and we tested it on the test set. Table 1 shows the configurations of the system we experimented, i.e. where the distance entailment engine applies *(i)* Word Overlap, *(ii)* cosine similarity, *(iii)* token edit distance, as the core distance algorithms. In every configuration, distance is calculated on lemmas, and a stopword list is defined to have no distance value between stopwords. We use the system off-the-shelf, applying some 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 basing on their syntactic structure.

		Train				**Test**			
	rel	*# pairs*	*Pr.*	*Rec.*	*Acc.*	*# pairs*	*Pr.*	*Rec.*	*Acc.*
WordOverlap	*yes*	100	0.71	0.52	**0.65**	50	0.72	0.52	**0.66**
	no	100	0.62	0.79		50	0.62	0.8	
Cosine sim.	*yes*	100	0.63	0.6	0.62	50	0.66	0.66	**0.66**
	no	100	0.62	0.64		50	0.66	0.66	
Token edit distance	*yes*	100	0.64	0.3	0.56	50	0.57	0.24	0.53
	no	100	0.54	0.83		50	0.51	0.82	

Table 1. EDITS performances on the Debatepedia data set using different configurations

Table 1 reports on the obtained results. Even using basic configurations of EDITS on a small data set, performances on Debatepedia test set are promising, and in line with performances of TE systems on RTE data sets (containing about 1000 pairs for training and 1000 for test). The results we obtained with EDITS best configuration (i.e. Word Overlap) are very close to the ones we obtained on a subset of the Debatepedia data set in [2], but in this new set of experiments the algorithm is not biased by the presence of a bigger percentage of a certain judgment during the training phase, since the data set is balanced.

As a second step of our evaluation phase, we consider the impact of EDITS performances on the acceptability of the arguments. We use admissibility-based semantics to identify the accepted arguments both on the correct argumentation frameworks of each Debatepedia topic (where entailment and contradiction relations are correctly assigned, i.e. the goldstandard), and on the frameworks generated assigning the relations resulted from the TE system (basing on its best configuration, i.e. Word Overlap). The precision of our overall framework in identifying the accepted arguments is on average 0.711 (i.e. arguments accepted by the framework and by the goldstandard w.r.t. a certain Debatepedia topic), and the recall is 0.72 (i.e. arguments accepted in the goldstandard and retrieved as accepted by the framework). Its accuracy (i.e. ability of the combined system to accept some arguments and discard some others) is 0.713, meaning that the TE system mistakes in relation assignment propagate in the BAF, but results are still satisfying and foster further research in this direction.

4. Conclusions

In this paper, we use textual entailment to analyze online dialogues in order to extract the abstract arguments and their relationships. This step is necessary to propose argu-

mentation theory as a suitable and natural way to support the users interacting within online discussions platforms. 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 relating them to each other. In particular, the TE system highlights two kinds of relations among the arguments: the entailment relation, when there is an inference between two arguments, and the attack relation when there is a contradiction among two arguments. We adopt bipolar argumentation frameworks [4,1] where the arguments either support or attack each others. The argumentation module returns the set of acceptable arguments. We evaluated our framework on a sample of topics extracted from Debatepedia, since it provided us with already annotated data to evaluate our system's performances. The accuracy of the overall framework in identifying the arguments (using TE) and correctly proposing the accepted arguments for each topic is about 71%.

Several research lines are considered to improve the proposed framework. We are using the TE module to reason about the modeling of the support relation in abstract argumentation theory. We are also considering to experiment our framework on scenarios different from Debatepedia, e.g. on Twitter, forums or media articles.

References

[1] G. Boella, D. M. Gabbay, L. W. N. van der Torre, and S. Villata. Support in abstract argumentation. In *Proceedings of COMMA, Frontiers in Artificial Intelligence and Applications vol. 216, IOS Press*, pages 111–122, 2010.

[2] E. Cabrio and S. Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of ACL*, 2012.

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

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

[5] C. I. Chesñevar and A.G. Maguitman. An argumentative approach to assessing natural language usage based on the web corpus. In *Proceedings of ECAI, IOS Press*, pages 581–585, 2004.

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

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

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

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

[10] M. Kouylekov and M. Negri. An open-source package for recognizing textual entailment. In *Proceedings of ACL 2010 System Demonstrations*, 2010.

[11] C. Reed and F. Grasso. Recent advances in computational models of natural argument. *Int. J. Intell. Syst.*, 22(1):1–15, 2007.

[12] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(4):983–, 2004.

[13] S. Buckingham Shum. Cohere: Towards Web 2.0 Argumentation. In *Proceedings of COMMA, Frontiers in Artificial Intelligence and Applications vol. 172, IOS Press*, pages 97–108, 2008.

[14] B. Verheij. Argumed - a template-based argument mediation system for lawyers and legal knowledge based systems. In *Proceedings of JURIX*, pages 113–130, 1998.

[15] A. Wyner and T. van Engers. A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *Proceedings of eGov*, 2010.