A Support Framework for Argumentative Discussions Management in the Web

Elena Cabrio, Serena Villata, and Fabien Gandon

INRIA Sophia Antipolis, France
{firstname.lastname}@inria.fr

Abstract. On the Web, wiki-like platforms allow users to provide arguments in favor or against issues proposed by other users. The increasing content of these platforms as well as the high number of revisions of the content through pros and cons arguments make it difficult for community managers to understand and manage these discussions. In this paper, we propose an automatic framework to support the management of argumentative discussions in wiki-like platforms. Our framework is composed by (i) a natural language module, which automatically detects the arguments in natural language returning the relations among them, and (ii) an argumentation module, which provides the overall view of the argumentative discussion under the form of a directed graph highlighting the accepted arguments. Experiments on the history of Wikipedia show the feasibility of our approach.

1 Introduction

On the Social Web, wiki-like platforms allow users to publicly publish their own arguments and opinions. Such arguments are not always accepted by other users on the Web, leading to the publication of additional arguments attacking or supporting the previously proposed ones. The most well known example of such kind of platform is Wikipedia[1] where users may change pieces of text written by other users to support, i.e., further specify them, or attack them, i.e., correcting factual errors or highlighting opposite points of view. Managing such kind of "discussions" using the revision history is a tricky task, and it may be affected by a number of drawbacks. First, the dimension of these discussions makes it difficult for both users and community managers to navigate, and more importantly, understand the meaning of the ongoing discussion. Second, the discussions risk to re-start when newcomers propose arguments which have already been proposed and addressed in the same context. Third, these discussions are not provided in a machine-readable format to be queried by community managers to discover insightful meta-information on the discussions themselves, e.g., discover the number of attacks against arguments about a particular politician concerning the economic growth during his government.

[1] http://en.wikipedia.org/wiki/Main_Page

In this paper, we answer the following research question: *how to support community managers in managing the discussions on the wiki pages?* This question breaks down into the following subquestions: (i) how to automatically discover the arguments and the relations among them?, and (ii) how to have the overall view of the ongoing discussion to detect the *winning* arguments? The answer to these sub-questions allows us to answer to further questions: how to detect repeated arguments and avoid loops of changes?, and how to discover further information on the discussion history? Approaches such as the lightweight vocabulary SIOC Argumentation [13] provide means to model argumentative discussions of social media sites, but they are not able to automatically acquire information about the argumentative structures. As underlined by Lange et al. [13], such a kind of automatic annotation needs the introduction of Natural Language Processing (NLP) techniques to automatically detect the arguments in the texts.

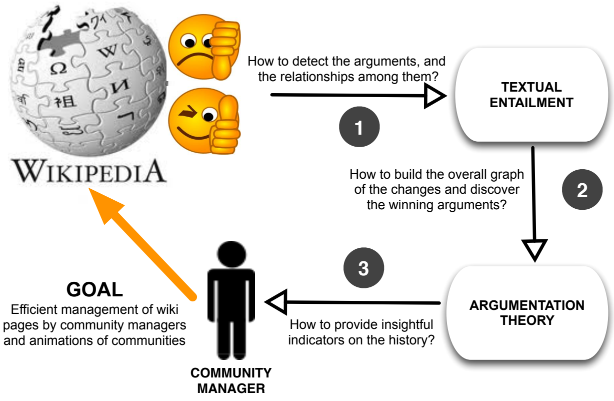

WIKIPEDIA
How to detect the arguments, and the relationships among them?
1
TEXTUAL ENTAILMENT
How to build the overall graph of the changes and discover the winning arguments?
2
3
ARGUMENTATION THEORY
How to provide insightful indicators on the history?
GOAL
Efficient management of wiki pages by community managers and animations of communities
COMMUNITY MANAGER

Fig. 1. An overview of the proposed approach to support community managers.

In this work, we propose a combined framework where a natural language module that automatically detects the arguments and their relations (i.e. *support* or *challenge*), is coupled with an argumentation module to have the overall view of the discussion and detect the winning arguments, as visualized in Figure 1.

First, to automatically detect natural language arguments and their relations, we rely on the Textual Entailment (TE) framework, proposed as an applied model to capture major semantic inference needs across applications in the NLP field [8]. Differently from formal approaches to semantic inference, in TE linguistic objects are mapped by means of semantic inferences at a textual level.

Second, we adopt abstract argumentation theory [9] to unify the results of the TE module into a unique argumentation framework able not only to provide the overall view of the discussion, but also to detect the set of *accepted* arguments relying on argumentation semantics. Argumentation theory aims at representing the different opinions of the users in a structured way to support decision making.

Finally, the generated argumentative discussions are described using an extension of the SIOC Argumentation vocabulary[2] thus providing a machine readable version. Such discussions expressed using RDF allow the extraction of a kind of "meta-information" by means of queries, e.g., in SPARQL. These meta-information cannot be easily detected by human users without the support of our automatic framework.

The aim of the proposed framework is twofold: on one side, we want to provide a support to community managers for notification and reporting, e.g., notify the users when their own arguments are attacked, and on the other hand, we support community managers to extract further insightful information from the argumentative discussions. As a case study, we apply and experiment our framework on Wikipedia revision history over a four-year period, focusing in particular on the top five most revised articles.

The paper is organized as follows. Section 2 provides some basic insights on abstract argumentation theory and textual entailment. Section 3 presents our combined framework to support the management of argumentative discussions in wiki-like platforms, and in Section 4 we report on the experimental setting and results. Section 5 presents and compares the related work.

2 Background: Argumentation and NLP

In this section, we provide notions of abstract argumentation theory and of textual entailment, essential to our work.

2.1 Abstract Argumentation Theory

A Dung-style argumentation framework [9] aims at representing conflicts among elements called *arguments* through a binary *attack* relation. It allows to reason about these conflicts in order to detect, starting by a set of arguments and the conflicts among them, which are the so called *accepted arguments*. The accepted arguments are those arguments which are considered as believable by an external evaluator, who has a full knowledge of the argumentation framework.

Definition 1 (Abstract argumentation framework AF **[9]).** *An abstract argumentation framework is a tuple* $\langle A, \rightarrow \rangle$ *where* A *is a finite set of elements called arguments and* $\rightarrow$ *is a binary relation called attack defined on* $A \times A$.

Dung [9] presents several acceptability semantics that produce zero, one, or several sets of accepted arguments. The set of accepted arguments of an argumentation framework consists of a set of arguments that does not contain an argument attacking another argument in the set. Roughly, an argument is *accepted* if all the arguments attacking it are rejected, and it is *rejected* if it has at least an argument attacking it which is accepted. In Figure 2.a, an example

[2] `http://rdfs.org/sioc/argument`

of abstract argumentation framework is shown. The arguments are visualized as circles, and the attack relation is visualized as edges in the graph. Gray arguments are the accepted ones. We have that argument a attacks argument b, and argument b attacks argument c. Using Dung's acceptability semantics [9], the set of accepted arguments of this argumentation framework is $\{a, c\}$.

The need of introducing also a positive relation among the arguments, i.e., a *support* relation, leads to the emergence of the so called *bipolar* argumentation frameworks [6].

Definition 2 (Bipolar argumentation framework BAF **[6]).** *A bipolar argumentation framework is a tuple* $\langle A, \rightarrow, \dashrightarrow \rangle$ *where* A *is a finite set of arguments,* $\rightarrow \subseteq A \times A$*, and* $\dashrightarrow$ *is a binary relation called support defined on* $A \times A$*.*

An example of bipolar argumentation framework is visualized in Figure 2.b where the dashed edge represents the support relation. For more details about acceptability semantics in BAFs, see [6].

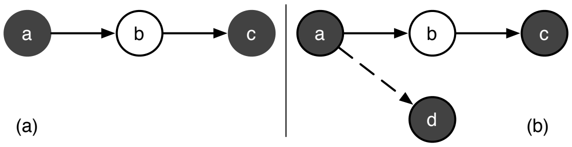


Fig. 2. Example of (a) an abstract argumentation framework, and (b) a BAF.

2.2 Textual Entailment

In the 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:

Example 1.
T: Jackson had three sisters: Rebbie, La Toya, and Janet, and six brothers: Jackie, Tito, Jermaine, Marlon, Brandon (Marlon's twin brother, who died shortly after birth) and Randy.
H: Jackson's siblings are Rebbie, Jackie, Tito, Jermaine, La Toya, Marlon, Randy and Janet.

Example 2 (Continued).
T: It was reported that Jackson had offered to buy the bones of Joseph Merrick (the elephant man) and although untrue, Jackson did not deny the story.
H: Later it was reported that Jackson bought the bones of The Elephant Man.

In Example 1, we can identify an inference relation between T and H (i.e. the meaning of H can be derived from the meaning of T), while in Example 2, T contradicts H. The notion of TE has been proposed [8] 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 using 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. 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 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.

3 The Combined Framework

In a recent work, Cabrio and Villata [2] propose to combine natural language techniques and Dung-like abstract argumentation to generate the arguments from natural language text and to evaluate this set of arguments to know which are the accepted ones, with the goal of supporting the participants in natural language debates (i.e. Debatepedia[4]). In particular, they adopt the TE approach, and in their experiments, they represent the TE relation extracted from natural language texts as a *support* relation in bipolar argumentation. In this paper, we start from their observations, and we apply the combined framework proposed in [2] to this new scenario.

Let us consider the argument in Example 3 from the Wikipedia article "United States", and its revised versions in the last four years[5]:

[3] *Dagan et al. (2009)* [8] provides an overview of the recent advances in TE.

[4] `http://bit.ly/Dabatepedia`

[5] Since we are aware that Wikipedia versions are revised daily, we have picked our example from a random dump per year. In Section 4.1, we provide more details about the Wikipedia sample we consider in our experiments.

Example 3.
In 2012: The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km2).
In 2011: The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).
In 2010: The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).
In 2009: The total land area of the contiguous United States is approximately 1.9 billion acres.

Several revisions have been carried out by different users during this four-year period, both to correct factual data concerning the U.S. surface, or to better specify them (e.g. providing the same value using alternative metric units). Following [2], we propose to take advantage of NLP techniques to automatically detect the relations among the revised versions of the same argument, to verify if the revisions done on the argument by a certain user at a certain point in time support the original argument (i.e. the user has rephrased the sentence to allow an easier comprehension of it, or has added more details), or attack it (i.e. the user has corrected some data, has deleted some details present in the previous version or has changed the semantics of the sentence providing a different viewpoint on the same content). Given the high similarities among the entailment and contradiction notions in TE and the support and attack relation in argumentation theory, we cast the described problem as a TE problem, where the T-H pair is a pair of revised arguments in two successive Wikipedia versions. We consider paraphrases as bidirectional entailment, and therefore to be annotated as a positive TE pair (i.e. support). Moreover, since the label *no entailment* includes both contradictions and pairs containing incomplete informational overlap (i.e. H is more informative than T), we consider both cases as *attacks*, since we want community managers to check the reliability of the corrected or deleted information. To build the T-H pairs required by the TE framework, for each argument we set the revised sentence as T and the original sentence as H, following the chronological sequence, since we want to verify if the more recent version entails or not the previous one, as shown in Example 4.

Example 4 (Continued).
pair id=70.1 entailment=NO
T (Wiki12): The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km2).
H (Wiki11): The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).

pair id=70.2 entailment=NO
T (Wiki11): The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).
H (Wiki10): The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).

pair id=70.3 entailment=YES
T (Wiki10): The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).
H (Wiki09): The total land area of the contiguous United States is approximately 1.9 billion acres.

On such pairs we apply a TE system, that automatically returns the set of arguments and the relations among them. The argumentation module starts from the couples of arguments provided by the TE module, and builds the complete argumentation framework involving such arguments. It is important to underline a main difference with respect to the approach of Cabrio and Villata [2]: here the argumentation frameworks resulting from the TE module represent a kind of *evolution* of the *same* argument during time in a specific Wikipedia article. From the argumentation point of view, we treat these arguments as separate instances of the same natural language argument giving them different names. Figure 3.a visualizes the argumentation framework of Example 4. This kind of representation of the natural language arguments and their evolution allows community managers to detect whether some arguments have been repeated in such a way that loops in the discussions can be avoided. The argumentation module, thus, is used here with a different aim from the previous approach [2]: it shows the *kind* of changes, i.e., positive and negative, that have been addressed on a particular argument, representing them using a graph-based structure.

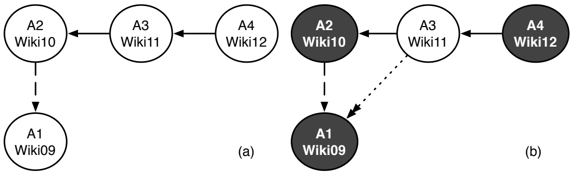


Fig. 3. The bipolar argumentation framework resulting from Example 4.

The use of argumentation theory to discover the set of winning, i.e., acceptable, arguments in the framework could seem pointless, since we could assume that winning arguments are only those arguments appearing in the most recent version of the wiki page. However, this is not always the case. The introduction of the support relation in abstract argumentation theory [6] leads to the introduction of a number of *additional attacks* which are due to the presence of an attack and a support involving the same arguments. The additional attacks introduced in the literature are visualized in Figure 4, where dotted double arrows represent the additional attacks. For the formal properties of these attacks and a comparison among them, see Cayrol and Lagasquie-Schiex [6].

The introduction of additional attacks is a key feature of our argumentation module. It allows us to support community managers in detecting further possible attacks or supports among the arguments. In particular, given the arguments and their relations, the argumentation module builds the complete framework

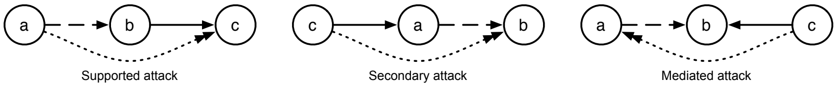


Fig. 4. The additional attacks arising due to the presence of a support relation.

adding the additional attacks, and computes the extensions of the bipolar framework. An example of such kind of computation is shown in Figure 3.b where an additional attack is introduced. In this example, the set of accepted arguments would have been the same with or without the additional attack, but there are situations in which additional attacks make a difference. This means that the explicit attacks put forward by the users on a particular argument can then result in *implicit* additional attacks or supports to other arguments in the framework. Consider the arguments of Example 5. The resulting argumentation framework (see Figure 5) shows that argument $A1$ (*Wiki09*) is implicitly supported by argument $A4$ (*Wiki12*) since the attack of $A4$ (*Wiki12*) against $A3$ (*Wiki11*) leads to the introduction of an additional attack against $A2$ (*Wiki10*). The presence of this additional attack reinstates argument $A1$ (*Wiki09*) previously attacked by $A2$ (*Wiki10*). The two accepted arguments at the end are $\{A1, A4\}$.

Example 5.
pair id=7.1 entailment=NO
T (Wiki12): In December 2007, the United States entered its longest post-World War II recession, prompting the Bush Administration to enact multiple economic programs intended to preserve the country's financial system.
H (Wiki11): In December 2007, the United States entered the longest post-World War II recession, which included a housing market correction, a subprime mortgage crisis, soaring oil prices, and a declining dollar value.

pair id=7.2 entailment=YES
T (Wiki11): In December 2007, the United States entered the longest post-World War II recession, which included a housing market correction, a subprime mortgage crisis, soaring oil prices, and a declining dollar value.
H (Wiki10): In December 2007, the United States entered its longest post-World War II recession.

pair id=7.3 entailment=NO
T (Wiki10): In December 2007, the United States entered its longest post-World War II recession.
H (Wiki09): In December 2007, the United States entered the second-longest post-World War II recession, and his administration took more direct control of the economy, enacting multiple economic stimulus packages.

Finally, in this paper we further enhance the framework proposed in [2] with a semantic machine readable representation of the argumentative discussions. We do not introduce yet another argumentation vocabulary, but we reuse the SIOC Argumentation module [13], focused on the fine-grained representation of discussions and argumentations in online communities.[6] The SIOC Argumentation model is grounded on DILIGENT [5] and IBIS[7] models. We extend the SIOC Argumentation vocabulary with two new properties `sioc_arg:challengesArg` and `sioc_arg:supportsArg` whose range and domain are `sioc_arg:Argument`. These properties represent challenges and supports from arguments to arguments, as required in abstract argumentation theory.[8] This needs to be done since in SIOC Argumentation challenges and supports are addressed from arguments towards `sioc_arg:Statement` only. Figure 6.a shows a sample of the semantic representation of Example 1 and 2 where *contradiction* is represented through `sioc_arg:challengesArg`, and *entailment* is represented through `sioc_arg:supportsArg`.

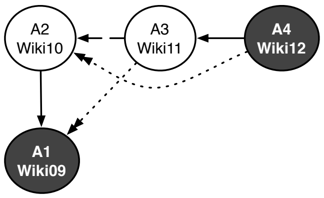
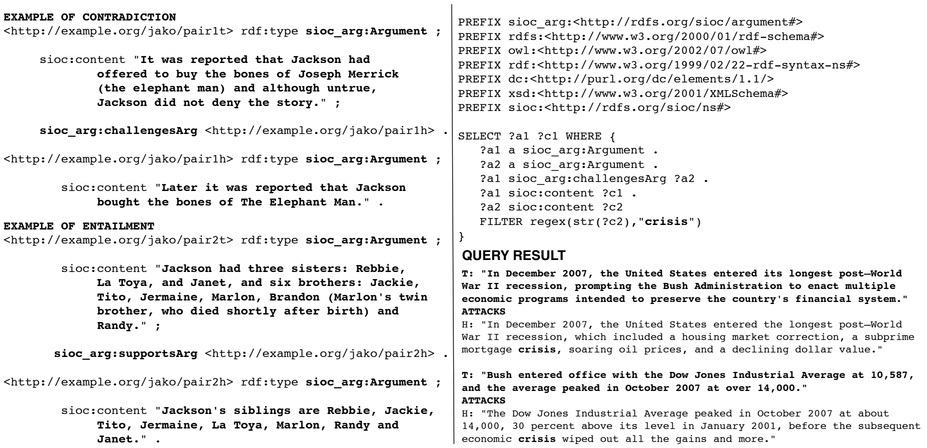


Fig. 5. The bipolar argumentation framework resulting from Example 5.



```
EXAMPLE OF CONTRADICTION
<http://example.org/jako/pair1t> rdf:type sioc_arg:Argument ;

     sioc:content "It was reported that Jackson had
             offered to buy the bones of Joseph Merrick
             (the elephant man) and although untrue,
             Jackson did not deny the story." ;

     sioc_arg:challengesArg <http://example.org/jako/pair1h> .

<http://example.org/jako/pair1h> rdf:type sioc_arg:Argument ;

        sioc:content "Later it was reported that Jackson
             bought the bones of The Elephant Man." .

EXAMPLE OF ENTAILMENT
<http://example.org/jako/pair2t> rdf:type sioc_arg:Argument ;

        sioc:content "Jackson had three sisters: Rebbie,
             La Toya, and Janet, and six brothers: Jackie,
             Tito, Jermaine, Marlon, Brandon (Marlon's twin
             brother, who died shortly after birth) and
             Randy." ;

       sioc_arg:supportsArg <http://example.org/jako/pair2h> .

<http://example.org/jako/pair2h> rdf:type sioc_arg:Argument ;

        sioc:content "Jackson's siblings are Rebbie, Jackie,
             Tito, Jermaine, La Toya, Marlon, Randy and
             Janet." .
```

```
PREFIX sioc_arg:<http://rdfs.org/sioc/argument#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX sioc:<http://rdfs.org/sioc/ns#>

SELECT ?a1 ?c1 WHERE {
   ?a1 a sioc_arg:Argument .
   ?a2 a sioc_arg:Argument .
   ?a1 sioc_arg:challengesArg ?a2 .
   ?a1 sioc:content ?c1 .
   ?a2 sioc:content ?c2
   FILTER regex(str(?c2),"crisis")
}
```

QUERY RESULT

```
T: "In December 2007, the United States entered its longest post—World
War II recession, prompting the Bush Administration to enact multiple
economic programs intended to preserve the country's financial system."
ATTACKS
H: "In December 2007, the United States entered the longest post—World
War II recession, which included a housing market correction, a subprime
mortgage crisis, soaring oil prices, and a declining dollar value."

T: "Bush entered office with the Dow Jones Industrial Average at 10,587,
and the average peaked in October 2007 at over 14,000."
ATTACKS
H: "The Dow Jones Industrial Average peaked in October 2007 at about
14,000, 30 percent above its level in January 2001, before the subsequent
economic crisis wiped out all the gains and more."
```

Fig. 6. (a) Sample of the discussions in RDF, (b) Example of SPARQL query.

The semantic version of the argumentative discussions can further be used by community managers to detect insightful meta-information about the discussions themselves. For instance, given the RDF data set being stored in a datastore with SPARQL endpoint, the community manager can raise a query

[6] For an overview of the argumentation models in the Social Semantic Web, see [15].

[7] `http://purl.org/ibis`

[8] The extended vocabulary can be downloaded at `http://bit.ly/SIOC_Argumentation`

like the one in Figure 6.b. This query retrieves all those arguments which attack another argument having in the content the word “crisis”. This simple example shows how the semantic annotation of argumentative discussions may be useful to discover in an automatic way those information which are difficult to be highlighted by a human user.

4 Experimental Setting

As a case study to experiment our framework we select the Wikipedia revision history. Section 4.1 describes the creation of the data set, Section 4.2 the TE system we used, while in Section 4.3 we report on obtained results.

4.1 Data Set

We create a data set to evaluate the use of TE to generate the arguments following the methodology detailed in [1]. We start from two dumps of the English Wikipedia (*Wiki 09* dated 6.03.2009, and *Wiki 10* dated 12.03.2010), and we focus on the five most revised pages[9] at that time (i.e. George W. Bush, United States, Michael Jackson, Britney Spears, and World War II). We then follow their yearly evolution up to now, considering how they have been revised in the next Wikipedia versions (*Wiki 11* dated 9.07.2011, and *Wiki 12* dated 6.12.2012).

After extracting plain text from the above mentioned pages, for both *Wiki 09* and *Wiki 10* each document has been sentence-splitted, and the sentences of the two versions have been automatically aligned to create pairs. Then, to measure the similarity between the sentences in each pair, following [1] we adopted the *Position Independent Word Error Rate (PER)*, i.e. a metric based on the calculation of the number of words which differ between a pair of sentences. For our task we extracted only pairs composed by sentences where major editing was carried out ($0.2 < PER < 0.6$), but still describe the same event.[10] For each pair of extracted sentences, we create the TE pairs setting the revised sentence (from *Wiki 10*) as T and the original sentence (from *Wiki 09*) as H. Starting from such pairs composed by the same revised argument, we checked in the more recent Wikipedia versions (i.e. *Wiki 11* and *Wiki 12*) if such arguments have been further modified. If that was the case, we created another T-H pair based on the same assumptions as before, i.e. setting the revised sentence as the T and the older sentence as the H (see Example 4). Such pairs have then been annotated with respect to the TE relation (i.e. *YES/NO entailment*), following the criteria defined and applied by the organizers of the Recognizing Textual Entailment Challenges (RTE)[11] for the two-way judgment task.

As a result of the first step (i.e. extraction of the revised arguments in *Wiki 09* and *Wiki 10*) we collected 280 T-H pairs, while after applying the procedure on the same arguments in *Wiki 11* and *Wiki 12* the total number of collected pairs is

[9] `http://bit.ly/WikipediaMostRevisedPages`

[10] A different extraction methodology has been proposed in [19].

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

452. To carry out our experiments, we randomly divided such pairs into training set (114 entailment, 114 no entailment pairs), and test set (101 entailment, 123 no entailment pairs). The pairs collected for the test set are provided in their unlabeled form as input to the TE system. To correctly train the TE system we balanced the data set with respect to the percentage of yes/no judgments. In Wikipedia, the actual distribution of attacks and supports among revisions of the same sentence is slightly unbalanced since generally users edit a sentence to add different information or correct it, with respect to a simple reformulation.[12]

To assess the validity of the annotation task and the reliability of the obtained data set, 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 140 argument pairs (randomly extracted). The statistical measure usually used in NLP to calculate the inter-rater agreement for categorical items is Cohen's kappa coefficient [4], that is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. More specifically, Cohen's kappa measures the agreement between two raters who each classifies N items into C mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. 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 (as defined by Pr(e)), $\kappa = 0$. For NLP tasks, the inter-annotator agreement is considered as significant when κ >0.6. Applying the formula (1) to our data, the inter-annotator agreement results in $\kappa = 0.7$. As a rule of thumb, this is a satisfactory agreement, therefore we consider these annotated data sets as the *goldstandard*[13], i.e. the reference data set to which the performances of our combined system are compared. As introduced before, the goldstandard pairs have then been further translated into RDF using SIOC Argumentation.[14]

4.2 TE System

To detect which kind of relation underlies each couple of arguments, we use the EDITS system (Edit Distance Textual Entailment Suite) version 3.0, an open-source software package for RTE[15] [12]. EDITS implements a distance-based framework which assumes that the probability of an entailment relation

[12] As introduced before, we set a threshold in our extraction procedure to filter out all the minor revisions, concerning typos or grammatical mistakes corrections.

[13] The dataset is available at `http://bit.ly/WikipediaDatasetXML`

[14] The obtained data set is downloadable at `http://bit.ly/WikipediaDatasetRDF`

[15] `http://edits.fbk.eu/`

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).[16] Within this framework the system implements different approaches to distance computation, i.e. both edit distance 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, that is then used at a test stage to assign a judgment and a confidence score to each test pair.

4.3 Evaluation

To evaluate our framework, we carry out a two-step evaluation: first, we assess the performances of EDITS to correctly assign the *entailment* and the *no entailment* relations to the pairs of arguments on the Wikipedia 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 Wikipedia training set to learn the model, and we test it on the test set. In the configurations of EDITS we experimented, the distance entailment engine applies *cosine similarity* and *word overlap* as the core distance algorithms. In both cases, distance is calculated on lemmas, and a stopword list is defined to have no distance value between stopwords.

Table 1. Systems performances on Wikipedia data set

		Train			**Test**		
EDITS configurations	*rel*	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*	*Accuracy*
WordOverlap	*yes*	0.83	0.82	**0.83**	0.83	0.82	**0.78**
	no	0.76	0.73		0.79	0.82	
CosineSimilarity	*yes*	0.58	0.89	0.63	0.52	0.87	0.58
	no	0.77	0.37		0.76	0.34	

Obtained results are reported in Table 1. Due to the specificity of our data set (i.e. it is composed by revisions of arguments), *word overlap* algorithm outperforms *cosine similarity* since there is high similarity between revised and original arguments (in most of the positive examples the two sentences are very close, or there is an almost perfect inclusion of H in T). For the same reason, obtained results are higher than in [2], and than the results obtained on average in RTE

[16] 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`

challenges. For these runs, we use the system off-the-shelf, applying its basic configuration. 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 based on their syntactic structure.

As a second step in our evaluation phase, we consider the impact of EDITS performances (obtained using word overlap, since it provided the best results) 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 [9] to identify the accepted arguments both on the correct argumentation frameworks of each Wikipedia revised argument (where entailment/contradiction relations are correctly assigned, i.e. the goldstandard), and on the frameworks generated assigning the relations resulted from the TE system judgments. The precision of the combined approach we propose in the identification of the accepted arguments is on average 0.90 (i.e. arguments accepted by the combined system and by the goldstandard w.r.t. a certain Wikipedia revised argument), and the recall is 0.92 (i.e. arguments accepted in the goldstandard and retrieved as accepted by the combined system). The F-measure (i.e. the harmonic mean of precision and recall) is 0.91, meaning that the TE system mistakes in relation assignment propagate in the argumentation framework, but results are still satisfying and foster further research in this direction. For this feasibility study, we use four Wikipedia versions, so the resulting AFs are generally composed by four couples of arguments connected by attacks or supports. Reduced AFs are produced when a certain argument is not revised in every Wikipedia version we considered, or when an argument is deleted in more recent versions. Using more revised versions will allow us to generate even more complex argumentation graphs.

5 Related Work

A few works investigate the use of Wikipedia revisions in NLP tasks. In Zanzotto and Pennacchiotti [19], two versions of Wikipedia and semi-supervised machine learning methods are used to extract large TE data sets, while Cabrio et al. [1] propose a methodology for the automatic acquisition of large scale context-rich entailment rules from Wikipedia revisions. [18] focus on using edit histories in Simple English Wikipedia to extract lexical simplifications. Nelken and Yamangil [17] compare different versions of the same document to collect users' editorial choices, for automated text correction and text summarization systems. Max and Wisniewski [14] create a corpus of natural rewritings (e.g. spelling corrections, reformulations) from French Wikipedia revisions. Dutrey et al. [10] analyze part of this corpus to define a typology of local modifications.

Other approaches couple NLP and argumentation. Chasnevar and Maguitman [7] use defeasible argumentation to assist the language usage assessment. Their system provides recommendations on language patterns and defeasible argumentation. No natural language techniques are applied 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. Differently from their work, 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 NL generation. Wyner and van Engers [16] 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. Beside the goals, several points distinguish the two works: *i)* their NLP module guides the user in writing the text using a restricted grammar and vocabulary, while we have no lexicon or grammar restrictions; *ii)* the inserted statements are associated with a mode indicating the relation between the existing and the input statements. We do not ask the user to explicit the relation among the arguments, we infer them using TE; *iii)* no evaluation of their framework is provided. Heras et al. [11] show how to model the opinions on business oriented websites using argumentation schemes. We share the same goal (i.e. providing a formal structure to on-line dialogues for evaluation,), but in our proposal we achieve it using an automatic technique to generate the arguments from natural language texts as well as their relations.

6 Conclusions

In this paper, we presented a framework to support community managers in managing argumentative discussions on wiki-like platforms. In particular, our approach proposes to automatically detect the natural language arguments and the relations among them, i.e., support or challenges, and then to organize the detected arguments in bipolar argumentation frameworks. This kind of representation helps community managers to understand the overall structure of the discussions and which are the winning arguments. Moreover, the generated data set is translated in RDF using an extension of the SIOC Argumentation vocabulary such that the discussions can be queried using SPARQL in order to discover further insightful information. The experimental evaluation shows that in 85% of the cases, the proposed approach correctly detects the accepted arguments.

SIOC[17] allows to connect the arguments to the users who propose them. This is important in online communities because it allows to evaluate the arguments depending on the expertise of their sources. In this paper, we do not represent users neither in the argumentation frameworks nor in the RDF representation of the discussions, and this is left as future work. Moreover, we plan to move from the crisp evaluation of the arguments' acceptability towards a more flexible evaluation where the expertise of the users proposing the arguments plays a role. As future work on the NLP side, we consider experimenting a TE system carrying out a three-way judgment task (i.e. *entailment*, *contradiction* and *unknown*), to

[17] `http://sioc-project.org`

allow for a finer-grained classification of non entailment pairs (i.e. to separate when T contradicts H, from when H is more informative than T).

References

1. E. Cabrio, B. Magnini, and A. Ivanova. Extracting context-rich entailment rules from wikipedia revision history. In *The People's Web Meets NLP Workshop*, 2012.
2. E. Cabrio and S. Villata. Natural language arguments: A combined approach. In *European Conference on Artificial Intelligence (ECAI)*, pages 205–210, 2012.
3. G. Carenini and J. D. Moore. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952, 2006.
4. J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.
5. A. G. Castro, A. Norena, A. Betancourt, and M. A. Ragan. Cognitive support for an argumentative structure during the ontology development process. In *Intl. Protege Conference*, 2006.
6. C. Cayrol and M.-C. Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. In S. Benferhat and J. Grant, editors, *Scalable Uncertainty Management*, volume 6929 of *LNCS*, pages 137–148. Springer Berlin Heidelberg, 2011.
7. C. I. Chesñevar and A. Maguitman. An argumentative approach to assessing natural language usage based on the web corpus. In *European Conference on Artificial Intelligence (ECAI)*, pages 581–585, 2004.
8. I. Dagan, B. Dolan, B. Magnini, and D. Roth. Recognizing textual entailment: Rational, evaluation and approaches. *JNLE*, 15(04):i–xvii, 2009.
9. P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.
10. C. Dutrey, H. Bouamor, D. Bernhard, and A. Max. Local modications and paraphrases in wikipedia's revision history. *SEPLN Journal*, 46:51–58, 2011.
11. S. Heras, K. Atkinson, V. J. Botti, F. Grasso, V. Julián, and P. McBurney. How argumentation can enhance dialogues in social networks. In *Computational Model of Arguments (COMMA)*, pages 267–274, 2010.
12. M. Kouylekov and M. Negri. An open-source package for recognizing textual entailment. In *ACL System Demonstrations*, pages 42–47, 2010.
13. C. Lange, U. Bojars, T. Groza, J. Breslin, and S. Handschuh. Expressing argumentative discussions in social media sites. In *SDoW*, 2008.
14. A. Max and G. Wisniewski. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *LREC*, 2010.
15. J. Schneider, T. Groza, and A. Passant. A review of argumentation for the social semantic web. *Semantic Web J.*, 2011.
16. A. Wyner and T. van Engers. A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *eGov*, 2010.
17. E. Yamangil and R. Nelken. Mining wikipedia revision histories for improving sentence compression. In *ACL (Short Papers)*, pages 137–140, 2008.
18. M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *HLT-NAACL*, pages 365–368, 2010.
19. F. Zanzotto and M. Pennacchiotti. Expanding textual entailment corpora from wikipedia using co-training. In *The People's Web Meets NLP Workshop*, 2010.