Inconsistencies Detection in Bipolar Entailment Graphs

Elena Cabrio, Serena Villata

Inconsistencies Detection in Bipolar Entailment Graphs

Elena Cabrio[1] **and Serena Villata**[2]
[2] CNRS, [1,2]University of Nice Sophia Antipolis, France
elena.cabrio@unice.fr; serena.villata@cnrs.fr

Abstract

English. In the latest years, a number of real world applications have underlined the need to move from Textual Entailment (TE) pairs to TE graphs where pairs are no more independent. Moving from single pairs to a graph has the advantage of providing an overall view of the issue discussed in the text, but this may lead to possible inconsistencies due to the combination of the TE pairs into a unique graph. In this paper, we adopt *argumentation theory* to support human annotators in detecting the possible sources of inconsistencies.

Italiano. *Negli ultimi anni, in svariate applicazioni sta sorgendo la necessità di passare da coppie di Textual Entailment (TE) a grafi di TE, in cui le coppie sono interconnesse. Il vantaggio dei grafi di TE è di fornire una visione globale del soggetto di cui si sta discutendo nel testo. Allo stesso tempo, questo può generare inconsistenze dovute all'integrazione di più coppie di TE in un unico grafo. In questo articolo, ci basiamo sulla teoria dell'argomentazione per supportare gli annotatori nell'individuare le possibili fonti di inconsistenze.*

1 Introduction

A Textual Entailment (TE) system (Dagan et al., 2009) automatically assigns to independent pairs of two textual fragments either an *entailment* or a *contradiction* relation. However, in some real world scenarios like analyzing costumer reviews about a service or product, these pairs cannot be considered as independent. For instance, all the reviews about a certain service need to be collected into a single graph, to understand the overall problems/merits of the service. The combination of TE pairs into a unique graph may generate *inconsistencies* due to the wrong relation assignment by the TE system, which could not have been identified if TE pairs were considered independently. The detection of such inconsistencies is usually left to human annotators, which later correct them. The need of processing such graphs to support annotators is therefore of crucial importance, particularly when dealing with big amounts of data. Our research question is *How to support annotators in detecting inconsistencies in TE graphs?*

The term *entailment graph* has been introduced by (Berant et al., 2010) as a structure to model entailment relations between propositional templates. Differently, in this paper we consider *bipolar entailment graphs* (BEGs), where two kinds of edges are considered, i.e., entailment and contradiction, to reason over the graph consistency.

We answer the research question by adopting *abstract argumentation theory* (Dung, 1995), a reasoning framework used to detect and solve inconsistencies in the so-called *argumentation graphs*, where nodes are called *arguments*, and edges represent a *conflict* relation. Argumentation semantics allows to compute *consistent* sets of arguments, given the conflicts among them.

We define the BEGincs (BEG-Inconsistencies) framework, which translates a BEG into an argumentation graph. It then provides to the annotators sets of arguments, following argumentation semantics, that are supposed to be consistent. If it is not the case, the TE system wrongly assigned some relations. Moving from single pairs to an overall graph allows for the detection of inconsistencies otherwise undiscovered. BEGincs does not identify the precise relation causing the inconsistency, but providing annotators with the consistent arguments sets, they are supported in narrowing the causes of inconsistency.

2 BEGincs framework

TE is a directional relation between two textual fragments. In various real world scenarios, these pairs cannot be considered as independent, and they need to be collected into a single graph. We define therefore a new framework involving *entailment graphs*, where pairs of textual fragments connected by semantic relations are also part of a graph that provides an overall view of the statements' interactions (*bipolar entailment graphs*).

Definition 1. *A bipolar entailment graph is a tuple* $BEG = \langle T, E, C \rangle$ *where* T *is a set of text fragments,* $E \subseteq T \times T$ *is an entailment relation between text fragments, and* $C \subseteq T \times T$ *is a contradiction relation between text fragments.*

This opens new challenges for TE, that originally considers the pairs as "self-contained" (i.e., the meaning of one text has to be derived from the meaning of the other). One challenge consists in checking BEGs to identify possible inconsistencies due to wrong relation assignments by the TE system. Figure 1 shows the architecture of the BEGincs framework to support human annotators in detecting inconsistencies in TE graphs.

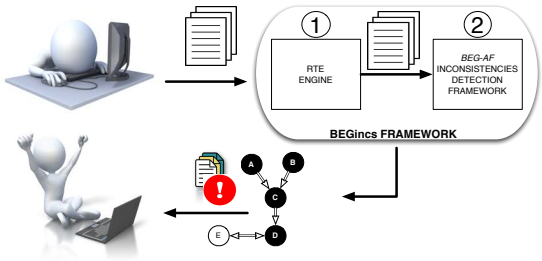

Figure 1: The BEGincs framework architecture.

Annotators provide the dataset to be checked as input of the BEGincs framework, which consists of two main modules: *(1)* a TE module, takes as input the dataset of text fragments, and returns the pairs annotated with the entailment or contradiction relations; and *(2)* a BEG-AF Inconsistencies Detection module, which translates the received BEGs into an argumentation framework such that argumentation semantics can be applied to retrieve consistent sets of arguments. The BEGincs framework returns through a user interface the starting BEGs highlighted with the consistent sets of text fragments. Checking them, annotators are able to detect errors in the annotation produced by the TE module (they will find inconsistent arguments in the returned sets), and correct the erroneous pairs.

2.1 Argumentation theory

An abstract argumentation framework (AF) (Dung, 1995) represents conflicts among elements called *arguments*. It is based on a binary *attack* relation among them, whose role is determined only by their relation with the other arguments. An AF encodes, through the *attack* relation, the existing conflicts within a set of arguments. It identifies then the conflict outcomes, i.e. which arguments should be accepted ("they survive the conflict") and which arguments should be rejected, according to some reasonable criterion. (Dung, 1995) presents several acceptability semantics that produce zero, one, or several *consistent* sets of accepted arguments. Such set of accepted arguments does not contain an argument conflicting with another argument in the set (*conflict free*). Following from this notion, an *admissible* set of arguments is required to be both internally coherent (*conflict-free*) and able to defend its elements. In BEGincs, we adopt admissibility based semantics. Roughly, an argument is accepted if all the arguments attacking it are rejected, and it is rejected if there is at least an argument attacking it which is accepted. The sets of accepted arguments computed using an acceptability semantics are called *extensions*, and the addition of another argument from outside the set will make it *inconsistent*.

2.2 Inconsistencies detection

To reuse abstract argumentation results and semantics for inconsistencies detection, we need to represent both the entailment and the contradiction relations of the bipolar entailment graph under the form of *attacks* between abstract arguments in an argumentation graph (Definition 2).

Definition 2. *A BEG-based argumentation framework is a tuple* $\langle A, \Rightarrow, \Leftrightarrow \rangle$ *where* A *is a set of text fragments called arguments,* $\Rightarrow$ *is a binary entailment relation on* A *(*$\Rightarrow \subseteq A \times A$*), and* $\Leftrightarrow$ *is a binary contradiction relation on* A *(*$\Leftrightarrow \subseteq A \times A$*). The set of arguments is* $\{a, b, \ldots \in A\}$.

BEG-AFs' consistent sets of arguments contain the text fragments that do not conflict with other fragments in the set (they are coherent). BEGincs uses the consistent sets of arguments computed following admissibility based argumentation semantics to support annotators in detecting inconsistencies. We need then to define the semantics of the entailment and contradiction relations in the

BEG-based argumentation framework (i.e. the behavior these relations have to satisfy in terms of conflict, since the only relation between arguments in abstract argumentation is the conflict relation).

Example 1.
T1: Natural gas vehicles run on natural gas, so emit significant amounts of greenhouse gases into the atmosphere, albeit smaller amounts than gasoline-fueled cars. To combat global warming, we should be focusing our energies and investments solely on 0-emission electric vehicles.
H: On the surface, natural gas cars seem alright, but the topic becomes a bit different when they are competing against zero emission alternatives (e.g. electric cars).

In Example 1, the text ($T1$) entails the hypothesis (H), i.e., $T1 \Rightarrow H$. Entailment is a directional relation (Dagan et al., 2009), that holds if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. In the pair, T is more specific than H (i.e., the more specific argument entails the more general one). In the argumentation setting, we have to reason over this feature to identify which constraints it poses in terms of conflicts among the text fragments. In particular, the following constraints emerge from the entailment relation: assuming *T entails H* holds, then *(i)* if there is a text fragment T_1 which contradicts H (negative TE) *then* T_1 contradicts also T ($T \equiv T_1$ does not entail $H \equiv T$), and *(ii)* if there is a text fragment T_2 which contradicts T *then* T_2 does not necessary contradict H too. These two constraints hold when a TE pair is inserted into an entailment graph. As a consequence, from the arguments' acceptance viewpoint: given that $T \Rightarrow H$, every time argument H is rejected, argument T is rejected too. We model the *entailment* relation such that, given that T entails H, T is accepted only if H is accepted too (Definit. 3)[1].

Definition 3. *Given a BEG-based argumentation framework $\langle A, \Rightarrow, \Leftrightarrow \rangle$, a translated BEG-based argumentation framework (BEG-AF) is a tuple $\langle \mathcal{A}, \longmapsto \rangle$ such that the set of arguments $\mathcal{A}$ is $\{a, b, \ldots \in A\} \cup \{X_{a,b}, Y_{a,b}, E_{a,b} \mid a, b \in A\}$, where $X_{a,b}, Y_{a,b}$ are the dummy arguments corresponding to the contradiction relation and $E_{a,b}$ is the dummy argument corresponding to the entailment relation, and $\longmapsto$ is a binary conflict relation over $\mathcal{A}$ such that: $b \longmapsto E_{a,b} \longmapsto a$ iff $a \Rightarrow b$.*

We have now to define the semantics of the *contradiction* relation (i.e., negative TE) in BEGs, see Example 2. (Marneffe et al., 2008) claims that contradiction occurs when two sentences *i)* are extremely unlikely to be true simultaneously, and *ii)* involve the same event. Starting from these considerations, the following constraint holds for the contradiction pairs: T and H conflict with each other (i.e. it is not possible to have both in a coherent and consistent set of arguments).

Example 2.
T2: Natural gas is the cleanest transportation fuel available today. If we want to immediately begin the process of significantly reducing greenhouse gas emissions, natural gas can help now. Other alternatives cannot be pursued as quickly.
H: On the surface, natural gas cars seem alright, but the topic becomes a bit different when they are competing against zero emission alternatives (e.g. electric cars).

Definition 4 models contradiction in BEG-AFs. The attack in (Dung, 1995) is directed from an argument to another argument while our contradiction leads to a cycle of attacks.

Definition 4. *Given a BEG-based argumentation framework $\langle A, \Rightarrow, \Leftrightarrow \rangle$, a BEG-AF is a tuple $\langle \mathcal{A}, \longmapsto \rangle$ such that $\mathcal{A}$ is the set of arguments, and $\longmapsto$ is a binary conflict relation over $\mathcal{A}$ such that:*
$a \longmapsto X_{a,b} \longmapsto Y_{a,b} \longmapsto b$, and
$b \longmapsto X_{b,a} \longmapsto Y_{b,a} \longmapsto a$, iff $a \Leftrightarrow b$.

Figure 2 summarizes the translation procedure, which is the core of our framework. We start with a BEG consisting of three text fragments (i.e., arguments A, B, C) from Ex. 1 and 2, where T1 is A, T2 is B, and H is C. The BEG is then translated into a BEG-AF where dummy arguments are introduced to express the semantics of the relations of entailment and contradiction, e.g., dummy argument $E_{A,C}$ represents the relation *A entails C* in the BEG-AF. The only relation allowed in a BEG-AF is the conflict relation $\longmapsto$. Therefore we have that a BEG-AF is a standard abstract AF, and we can apply admissibility based argumentation semantics to retrieve consistent sets of arguments. Acceptability semantics return the extension of the BEG-AF (i.e., the black nodes in Fig. 2), where arguments C,A are accepted, and dummy arguments are filtered out from the set of accepted ones.

We prove now that our BEG-AF actually satisfies the semantics of the entailment relation.

[1] See (Cabrio and Villata, 2013) for a comparison of the *entailment* wrt the *support* relation (Boella et al., 2010).

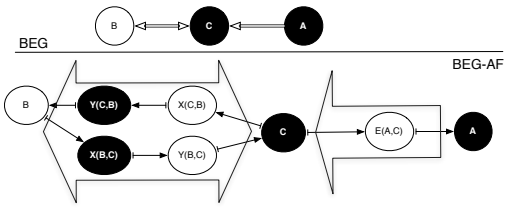

Figure 2: Translation from a BEG to a BEG-AF.

Proposition 1 (Semantics of *entailment*)**.** *Given a BEG-AF, if it holds that $T \Rightarrow H$ and text fragment T is accepted, then fragment H is accepted too.*

Proof. We prove the contrapositive. If it holds that $T \Rightarrow H$ and text fragment H is not accepted, then text fragment T is not accepted. Assume that $T \Rightarrow H$ and assume that argument H is not accepted, then dummy argument $E_{T,H}$ is accepted. Consequently, T is not accepted, i.e., rejected. □

We need to add two nodes, i.e., dummy arguments $X_{a,b}$ and $Y_{a,b}$, to represent a contradiction while we only need one node, i.e., dummy argument $E_{a,b}$, to represent entailment, since preserving the semantics of a contradiction holding between two text fragments means that the two text fragments cannot be together in a consistent set of arguments. To avoid the two being both accepted, we need to introduce two dummy arguments so that: a (accepted) $\longmapsto$ $X_{a,b}$ (rejected), $X_{a,b} \longmapsto Y_{a,b}$ (accepted), and $Y_{a,b} \longmapsto b$ (rejected). In this way, if a is accepted then b is rejected, and viceversa. A unique dummy argument between a and b would not ensure such behavior.

Existing works combine NLP and argumentation theory, e.g. (Chesñevar and Maguitman, 2004; Carenini and Moore, 2006; Wyner and van Engers, 2010; Feng and Hirst, 2011) with different purposes. However, only our previous work (Cabrio and Villata, 2012) combines TE with AF, but here our goal is to introduce a framework for inconsistencies detection in TE annotations.

3 Experimental setting

Data set. We added 60 pairs to the Debatepedia dataset [2] (extracted from a sample of Debatepedia[3] debates (Cabrio and Villata, 2012)), resulting in 160 pairs as training set, and 100 pairs as test set (balanced wrt to entailment/contradiction).

Evaluation. *First step*: we assess the performances of the TE system to correctly assign the TE relations to the pairs of arguments in the dataset. *Second step*: we evaluate how much such performances impact on the flattening of the BEG-AF, i.e., how much a wrong assignment of a relation to a pair of arguments is propagated in the AF. It is actually to detect such wrong assignments that the BEGincs framework has been conceived.

To recognize TE, we tested several algorithms from the EOP[4], i.e. BIUTEE (Stern and Dagan, 2011), TIE[5] and EDITS (Kouylekov and Negri, 2010). BIUTEE obtained the best results on Debatepedia (configuration exploiting all available knowledge resources): Acc:0.71, Rec:0.94, Pr:0.66, F-meas:0.78. As baseline we use a token-based version of the Levenshtein distance algorithm, i.e. EditDistanceEDA in the EOP (Acc:0.58, Rec:0.61, Pr:0.59, F-meas:0.59).

Then, we consider the impact of the best TE configuration on the arguments acceptability. We use admissibility-based semantics to identify the accepted arguments both on *i)* the goldstandard entailment graphs of Debatepedia topics, and *ii)* on the graphs generated using the relations assigned by BIUTEE. On the 10 Debatepedia graphs, BEGincs avg pr:0.68, avg rec:0.91, F-meas:0.77. BIUTEE mistakes in relation assignment propagate in the AF, but results are promising. The incons. detection module takes ∼1 sec. to analyze a BEG of 100 nodes and 150 relations.

4 Concluding remarks

We have presented BEGincs, a new formal framework that, translating a BEG into an argumentation graph, returns inconsistent set of arguments, if a wrong relation assignment by the TE system occurred. These inconsistent arguments sets are then used by annotators to detect the presence of a wrong assignment, and if so, to narrow the set of possibly erroneous relations. If no mistakes are produced in relation assignment, by definition BEGincs semantics return consistent arguments sets.

Assuming that in several real world scenarios TE pairs are interconnected, we ask to the NLP community to contribute in the effort of building suitable resources. In BEGincs, we plan to verify and ensure transitivity of BEGs.

[2] The only available dataset of T-H pairs combined into bipolar entailment graphs.

[3] http://idebate.org/

[4] http://bit.ly/ExcitementOpenPlatform

[5] http://bit.ly/MaxEntClassificationEDA

References

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

G. Boella, D. M. Gabbay, L. W. N. van der Torre, and S. Villata. 2010. Support in abstract argumentation. In P. Baroni, F. Cerutti, M. Giacomin, and G. R. Simari, editors, *COMMA*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 111–122. IOS Press.

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

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

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

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

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

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

V. Wei Feng and G. Hirst. 2011. Classifying arguments by scheme. In *Procs of ACL-2012*, pages 987–996.

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

M.C. De Marneffe, A.N. Rafferty, and C.D. Manning. 2008. Finding contradictions in text. In *Procs of ACL*.

A. Stern and I. Dagan. 2011. A confidence model for syntactically-motivated entailment proofs. In *Proceedings of RANLP 2011*.

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