Semantic Annotation of Textual Entailment

Assaf Toledo
a.toledo@uu.nl

Stavroula Alexandropoulou
s.alexandropoulou@uu.nl

Sophia Katrenko
sophia@katrenko.com

Heidi Klockmann
h.e.Klockmann@uu.nl

Pepijn Kokke
pepijn.kokke@gmail.com

Yoad Winter
y.winter@uu.nl

Utrecht University

1 Abstract

We introduce a new formal semantic model for annotating textual entailments, that describes restrictive, intersective and appositive modification. The model contains a formally defined interpreted lexicon, which specifies the inventory of symbols and the supported semantic operators, and an informally defined annotation scheme that instructs annotators in which way to bind words and constructions from a given pair of premise and hypothesis to the interpreted lexicon. We explore the applicability of the proposed model to the Recognizing Textual Entailment (RTE) 1-4 corpora and describe a first-stage annotation scheme based on which manual annotation work was carried out. The constructions we annotated were found to occur in 80.65% of the entailments in RTE 1-4 and were annotated with cross-annotator agreement of 68% on average. The annotated RTE corpora are publicly available for the research community.

2 Introduction

The RTE challenges (Dagan et al., 2006) aim to automatically determine whether an entailment relation obtains between a naturally occurring **text** sentence (T) and a **hypothesis** sentence (H). The RTE corpus (Bar Haim et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009), which is currently the only available resource of textual entailments, marks entailment candidates as valid/invalid.[1]

Example 1

- T: The head of the Italian opposition, Romano Prodi, was the last president of the EC.
- H: Romano Prodi is a former president of the EC.[2]
- Entailment: Valid

This categorization contains no indication of the linguistic processes that underlie entailment. In the lack of a gold standard of inferential phenomena, entailment systems can be compared based on their performance, but their inferential processes are not directly accessible for analysis.

The goal of this work is to elucidate some central inferential processes underlying entailments in the RTE corpus. By doing that, we aim to advance the possibility of creating a benchmark for modeling entailment recognition. We presume that this goal is to be achieved incrementally by modeling increasingly complex semantic phenomena. To this end, we employ a standard model-theoretic approach to entailment in order to combine gold standard annotations with a computational framework. The model

[1] Pairs of sentences in RTE 1-3 are categorized in two classes: *yes-* or *no-entailment*; pairs in RTE 4-5 are categorized in three classes: *entailment*, *contradiction* and *unknown*. We label the judgments *yes-entailment* from RTE 1-3 and *entailment* from RTE 4-5 as *valid*, and the other judgments as *invalid*.

[2] Pair 410 from the test set of RTE 2. *EC* stands for European Commission

contains a formally defined interpreted lexicon, which specifies the inventory of symbols and semantic operators that are supported, and an informally defined annotation scheme that instructs annotators how to bind words and constructions from a given T-H pair to entries in the interpreted lexicon. Our choice to focus on the semantic phenomena of restrictive, intersective and appositive modification is driven by their predominance in the RTE datasets, the ability to annotate them with high consistency and the possibility to capture their various syntactic expressions by a limited set of concepts.

However, currently we are only at the first stages of implementing the theoretical semantic model using an annotation platform combined with a theorem prover. In the course of the development of this model we adopted a narrower annotation scheme by which modification phenomena were annotated in all valid entailment pairs from RTE 1-4 without accounting for the way in which the annotated phenomena contribute to the inference being made. This work allowed us to perform data analysis and to further learn about the phenomena of interest as part of the development of the semantic model.

The structure of this paper is as follows. Section 3 reviews some related methods used in Bos et al. (2004) and MacCartney and Manning (2007). In Section 4 we introduce the formal semantic model on which we rely and use it for analyzing some illustrative textual entailments. Section 5 points out a challenge in applying this model to parts of the RTE data and describes our first-stage annotation scheme. We elaborate on the methods employed in applying this scheme to the datasets of RTE 1-4, and present some quantitative data on the targeted phenomena and inter-annotator agreement. Section 6 concludes.

3 Related Work

Bos and Markert (2005) utilizes a CCG parser (Bos et al., 2004) to represent the text and hypothesis in discourse representation structures (DRSs, Kamp and Reyle 1993) that encapsulate information on argument structure, polarity, etc. The DRSs of the text and hypothesis are then translated into formulae in first order logic, and a theorem prover is used in order to search whether there is a logical proof from the text formula to the hypothesis formula. The system reached a relatively high precision score of 76% in recognizing the positive cases in RTE 2 but suffered from a very low recall of 5.8%.

MacCartney and Manning (2007)'s system recognizes monotonic relations (or lack thereof) between aligned lexical items in the text and hypothesis and employs a model of compositional semantics to calculate a sentence-level entailment prediction. The recognition of monotonic relations is done using an adapted version of Sanchez Valencia's Natural Logic (Valencia, 1991), the alignment between the text and hypothesis is based on a cost function that extends Levenshtein string-edit algorithm, and the entailment is classified by a decision tree classifier, trained on a small data set of 69 handmade problems. The system was tested on RTE 3 and achieved relatively high precision scores of 76.39% and 68.06% on the positive cases in the development and test sets respectively. This system also suffers from low recall scores of 26.70% and 31.71% respectively.

The model we propose in this work diverges from these systems in two respects: (a) its first goal is to develop gold standard semantic annotations based on a general formal semantic model; (b) it does not aim to represent phenomena that are not accounted for in this model. For example, consider the following inference, which is based on causal reasoning: *Khan sold nuclear plans* $\Rightarrow$ *Khan possessed nuclear plans*.[3] Causal reasoning and lexical relations are not part of the semantic phenomena addressed in this paper, and a pattern in the form of *X sold Y* $\Rightarrow$ *X possessed Y* should be defined ad-hoc by annotators to align the instances of the verbs *sell* and *possess*. This approach allows us to concentrate on the logical aspects of textual entailment, while phenomena involving lexical semantics and world knowledge are handled by a shallow analysis.[4]

[3]This example of causal reasoning is taken from MacCartney and Manning (2007).

[4]Another related work, which approaches inference in natural language as part of a semantic paradigm, is the FraCaS test suite (Cooper et al., 1996). This suit concerns examples that mainly rely on generalized quantification, argument monotonicity, plurality, anaphora resolution, ellipsis, etc. Entailments based on these phenomena are not very common in the RTE data that are analyzed here. Further research is needed in order to integrate data like those in FraCaS into a formal annotation scheme like the one suggested in this paper.

4 Theoretical background and RTE examples

To model entailment in natural language, we assume that entailment describes a *preorder* on natural language sentences. Thus, we assume that any sentence trivially entails itself (reflexivity); and given two entailments $T_1 \Rightarrow H_1$ and $T_2 \Rightarrow H_2$ where H_1 and T_2 are identical sentences, we assume $T_1 \Rightarrow H_2$ (transitivity). A computational theory of entailment should describe an approximation of this preorder on natural language sentences. We use a standard model-theoretical extensional semantics, based on the simple *partial order* on the domain of *truth-values*. Each model M assigns sentences a truth-value in the set $\{0, 1\}$. Such a Tarskian theory of entailment is considered adequate if the intuitive entailment preorder on sentences can be described as the pairs of sentences T and H whose truth-values $[\![T]\!]^M$ and $[\![H]\!]^M$ satisfy $[\![T]\!]^M \leq [\![H]\!]^M$ for all models M. In this section we give the essentials of this model-theoretic approach to entailment that are relevant to the annotated phenomena and illustrate it using a small interpreted lexicon, simplifying the analysis of some representative examples from the RTE.

4.1 An interpreted lexicon

The interpreted lexicon presented in Table 1 illustrates our treatment of major lexical categories over types e, t and their functional compounds. Our aim is to allow binding of words and expressions in entailment data to the lexicon. Each word is stated in its literal form, the type assigned to it, and its denotation in intended models. Denotations that are assumed to be arbitrary in intended models are given in boldface. For example, the intransitive use of the verb *sit* is assigned the type et and its denotation $\mathbf{sit}$ is an arbitrary function of this type. By contrast, other lexical items have their denotations restricted by the intended models. For example, the definite article *the* is assigned the type $(et)e$ and its denotation is fixed as the *iota* operator. The functions that we use for defining denotations are specified in Figure 1. Several items in the lexicon are assigned more than one type and/or more than one denotation due to ambiguity in natural language. The following list explains some of the main items in the lexicon:

- The coordinator *and*, when appearing as predicate conjunction, is analyzed as a function - AND, mapping any two et predicates A and B to a predicate that sends every entity e to the truth-value of the conjunction $A(x) \wedge B(x)$.
- The copular *is* and the article *a* in copular sentences (e.g. *Dan is a man* / *Dan is short*) are analyzed as identity functions IS and A of type $(et)(et)$ respectively. In copula sentences that express an equality relation (e.g. *Dan is Jan*), *is* is analyzed by the equality function IS_{eq} of type $e(et)$.
- The word *some* denotes the existential quantifier SOME, as it is used in intransitive sentences such as *some man sat* (transitive sentences like *Jan saw some man* are not treated here).
- The relative pronoun *who* allows noun modification either by a restrictive relative clause denoted by WHO_R or by an appositive clause denoted by WHO_A. WHO_R is expressed in sentences such as *the alien who is a nun sat*, in which the pronoun creates a complex predicate, *alien who is a nun*. WHO_A appears in sentences such as *the alien, who is a nun, sat* where the pronoun adds information on a given entity x. The resulting entity is x if A holds of x, and undefined otherwise.
- The adjectives *short* and *Dutch*, when appearing as modifiers, restrict the denotation of the noun they attach to: *a short/Dutch man* is *a man*. *Dutch* is furthermore *intersective*: *a Dutch man* is invariably *Dutch*. The predicate *Dutch* is defined as an arbitrary constant $\mathbf{dutch}$ of type et. The modifier is derived by a function I_m identical to AND. The restrictive modifier *short* is defined by the function R_m and a constant **short** of type $(et)(et)$. The predicative denotation of *short* is defined using the function P_r as the set of "short things" - by applying the constant to all entities.

See Pratt-Hartmann and Moss (2009) for a wider coverage of some of the same semantic ground that goes further in dealing with comparative constructions and transitive verbs.

4.2 Analyzing entailments using the interpreted lexicon

Some central logical semantic aspects of entailments from the RTE can be formally analyzed using the lexicon in Table 1. We analyze entailments by binding expressions in the RTE data to structurally equivalent expressions containing items in the interpreted lexicon. This analysis is three-fold:

Word	Type	Denotation	Remarks
Dan, Jan, Vim	e	$\mathbf{dan}$, $\mathbf{jan}$, $\mathbf{vim}$	proper name
man, nun, alien	et	$\mathbf{man}$, $\mathbf{nun}$, $\mathbf{alien}$	intrans. noun
sat	et	$\mathbf{sit}$	intrans. verb
saw	$e(et)$	$\mathbf{see}$	trans. verb
and	$(et)((et)(et))$	AND	pred. conj. (coordinator)
is	$(et)(et)$	IS	copula (modifier)
is	$e(et)$	IS_{eq}	copula (equality)
a	$(et)(et)$	A	indef. article (modifier)
the	$(et)e$	THE	def. article (iota)
some	$(et)((et)t)$	SOME	indef. determiner
who	$(et)((et)(et))$	WHO_R	res. rel. pronoun (coordinator)
who	$(et)(ee)$	WHO_A	app. rel. pronoun
Dutch, black	et	$\mathbf{dutch}_{et}$, $\mathbf{black}_{et}$	int. adjective (predicate)
Dutch, black	$(et)(et)$	$I_m(\mathbf{dutch}_{et})$, $I_m(\mathbf{black}_{et})$	int. adjective (modifier)
short	et	$P_r(\mathbf{short}_{(et)(et)})$	res. adjective (predicate)
short	$(et)(et)$	$R_m(\mathbf{short}_{(et)(et)})$	res. adjective (modifier)
slowly	$(et)(et)$	$R_m(\mathbf{slowly}_{(et)(et)})$	res. adverb (modifier)

Table 1: An Interpreted Lexicon

$$\text{AND} = \lambda A_{et}.\lambda B_{et}.\lambda x_e.B(x) \wedge A(x)$$

$$\text{IS} = \lambda A_{et}.A$$

$$\text{IS}_{eq} = \lambda x_e.\lambda y_e.x = y$$

$$\text{A} = \text{IS} = \lambda A_{et}.A$$

$$\text{THE} = \iota_{(et)e} = \lambda A_{et}.\begin{cases} a & \text{if } A = (\lambda x_e.x = a) \\ \text{undefined} & \text{otherwise} \end{cases} \quad \text{(iota operator)}$$

$$\text{SOME} = \lambda A_{et}.\lambda B_{et}.\exists x_e.A(x) \wedge B(x)$$

$$\text{WHO}_R = \text{AND} = \lambda A_{et}.\lambda B_{et}.\lambda x_e.B(x) \wedge A(x)$$

$$\text{WHO}_A = \lambda A_{et}.\lambda x_e.\iota(\lambda y.y = x \wedge A(x))$$

$$P_r = \lambda M_{(et)(et)}.\lambda x_e.M(\lambda y_e.1)(x) \quad \text{deriving a predicate from a general modifier}$$

$$I_m = \text{AND} = \lambda A_{et}.\lambda B_{et}.\lambda x_e.B(x) \wedge A(x) \quad \text{deriving an intersective modifier}$$

$$R_m = \lambda M_{(et)(et)}.\lambda A_{et}.\lambda x_e.M(A)(x) \wedge A(x) \quad \text{deriving a restrictive modifier}$$

Figure 1: Functions used in the interpreted lexicon

1. Phenomena Simplification: we simplify the text and hypothesis to exclude inferential phenomena that we do not handle in the scope of this work. For instance, in Example 2, the inference *Google operates on the web* ⇒ *Google is on the web* is based on lexical knowledge, which we do not address here, and therefore it is handled as part of the simplification step.
2. Binding to Lexicon: we bind the constructions in the data to parallel constructions in the interpreted lexicon that share the same structure and semantic properties. This step produces a text $\text{T}_{Lexicon}$ and a hypothesis $\text{H}_{Lexicon}$ as new structurally equivalent versions of the simplified text and hypothesis. The parse trees are assumed in a way that allows to apply the interpreted lexicon.
3. Proof of Entailment: using predicate calculus and lambda calculus reductions, we establish a logical proof between $\text{T}_{Lexicon}$ and $\text{H}_{Lexicon}$.[5]

Example 2

- Data:
  - T: The largest search engine on the web, Google receives over 200 million queries each day through its various services.

[5]The only higher-order constants in the above lexicon are the $(et)(et)$ constants attributed to non-intersective restrictive modifiers. Treating them in predicate calculus theorem provers may require some *ad hoc* assumptions.

– H: Google operates on the web.[6]

1. Phenomena Simplification:
   In the text: adding an overt appositive WH pronoun to match the interpreted lexicon:
   - $\text{T}_{Original}$: The largest search engine on the web, Google receives...
   - T_{Simple}: The largest search engine on the web, which is Google, receives...

   In the hypothesis: reducing the meaning of 'X operates on Y' to 'X is on Y':
   - $\text{H}_{Original}$: Google operates on the web
   - H_{Simple}: Google is on the web
2. Binding to Lexicon:
   Text[7,8]:
   - T_{Simple}: [The largest search engine on the web, which is Google,] receives...
   - $\text{T}_{Lexicon}$: [The short Dutch man, who is Jan,] saw Dan

   Hypothesis:
   - H_{Simple}: Google [is [on the web]]
   - $\text{H}_{Lexicon}$: Jan [is Dutch]
3. Proof of entailment $\text{T}_{Lexicon} \Rightarrow \text{H}_{Lexicon}$: Let M be an intended model,
   $[\![$ [[*The* [*short Dutch man*]], [*who* [*is Jan*]],] *saw Dan* $]\!]^M$

$= (\mathbf{see}(\mathbf{dan}))((\text{WHO}_A(\text{IS}_{eq}(\mathbf{jan})))(\iota((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man})))))$	analysis
$= (\mathbf{see}(\mathbf{dan}))(\iota(\lambda y.y = (\iota((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man}))) \wedge \mathbf{jan} = (\iota((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man}))))))$	def. of WHO_A + IS_{eq}, func application
By definition of ι: $\mathbf{jan} = \iota((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man})))$	
$\Rightarrow \mathbf{jan} = \iota(\lambda y_e.(\mathbf{short}(\lambda x_e.\mathbf{man}(x) \wedge \mathbf{dutch}(x)))(y) \wedge \mathbf{man}(y) \wedge \mathbf{dutch}(y))$	def. of R_m, I_m, $\wedge$ + func. application
By definition of ι: $(\mathbf{short}(\lambda x_e.\mathbf{man}(x) \wedge \mathbf{dutch}(x)))(\mathbf{jan}) \wedge \mathbf{man}(\mathbf{jan}) \wedge \mathbf{dutch}(\mathbf{jan})$	
$\leq \mathbf{dutch}(\mathbf{jan}) = (\text{IS}(\mathbf{dutch}))(\mathbf{jan}) = [\![$*Jan [is Dutch]*$]\!]^M$	def. of $\wedge$, IS + analysis

A crucial step in this analysis is our assumption that *on the web* is an intersective modifier of *search engine*. This allows the subsumption of *search engine on the web* by *on the web*. In the interpreted lexicon we describe this behavior using the intersective denotation of the modifier *Dutch*. Let us investigate further the implications of this annotation in the following hypothetical example.

Example 3

1. Pair 1: T_1: Jan is a short Dutch man $\not\Rightarrow$ H_1: Jan is a short man — no entailment
2. Pair 2: T_2: Jan is a black Dutch man $\Rightarrow$ H_2: Jan is a black man — entailment

From a purely textual/syntactic point of view, these two T-H pairs are indistinguishable. The lexical overlap between the text and hypothesis in both pairs is 100%. This does not allow entailment systems to rely on textual measurements to identify that the pairs need to be classified differently. Such a perfect score of overlap may lead to a false positive classification in Pair 1 or conversely, to a false negative in Pair 2. Also syntactically, both *short* and *black* serve as adjectives attached to a noun phrase - *Dutch man*. There is nothing in this syntactic configuration to suggest that omitting *Dutch* in Pair 1 might result in a different entailment classification than omitting it in Pair 2. However, from a semantic point of view, based on annotations of abstract relations between predicates and their modifiers, we can correctly analyze both the non-validity of the entailment in Pair 1 and the validity of the entailment in Pair 2.

- Analysis of Pair 1

  To validate that there is no entailment between a text and a hypothesis means to show that there is an intended model $M = \langle E, I \rangle$ in which there is no $\leq$ relation between their denotations.

[6] Pair 955 from the test set of RTE 4 (Giampiccolo et al., 2008).

[7] Note that the post-nominal intersective modifier *on the web* is bound to a pre-nominal modifier *Dutch*. This is done in order to match the vocabulary of the interpreted lexicon, in which the only intersective modifier is *Dutch*.

[8] In this example, T_{Simple} (consequently from $T_{Original}$) is structurally ambiguous between *The* [*largest* [*search engine on the web*]], *which is Google, receives...* and *The* [[*largest search engine*] *on the web*], *which is Google, receives....* We illustrate the former analysis here. The latter analysis can be handled in a similar vein.

Let M be an intended model that satisfies the following:

- $\mathbf{man}_{et}$ characterizes $\{\mathbf{dan}, \mathbf{jan}, \mathbf{vim}\}$
- $\mathbf{dutch}_{et}$ characterizes $\{\mathbf{jan}, \mathbf{vim}\}$
- $\mathbf{short}(\mathbf{man})_{et}$ characterizes $\{\mathbf{dan}\}$
- $\mathbf{short}(\lambda y_e.\mathbf{man}(y) \wedge \mathbf{dutch}(y))_{et}$ characterizes $\{\mathbf{jan}\}$

Let us assume parse trees as follows:

- Text: *Jan* [*is* [*a* [*short* [*Dutch man*]]]]
- Hypothesis: *Jan* [*is* [*a* [*short man*]]]

Consider the denotations of the text and hypothesis in the model M:

- Text:

  $[\![$ *Jan* [*is* [*a* [*short* [*Dutch man*]]]] $]\!]^M$

= $(\text{IS}(\text{A}((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man})))))(\mathbf{jan})$	analysis
= $((R_m(\mathbf{short}))((I_m(\mathbf{dutch}))(\mathbf{man})))(\mathbf{jan})$	def. of A, IS
= $(((\lambda M_{(et)(et)}.\lambda A_{et}.\lambda y_e.M(A)(y) \wedge A(y))(\mathbf{short}))$ $(((\lambda A_{et}.\lambda B_{et}.\lambda x_e.B(x) \wedge A(x))\ (\mathbf{dutch}))(\mathbf{man})))(\mathbf{jan})$	def. of I_m, R_m
= $1 \wedge 1 \wedge 1 = 1$	func. application + denotations in M

- Hypothesis:

  $[\![$ *Jan* [*is* [*a* [*short man*]]] $]\!]^M$

= $(\text{IS}(\text{A}((R_m(\mathbf{short}))(\mathbf{man}))))(\mathbf{jan})$	analysis
= $((R_m(\mathbf{short}))(\mathbf{man}))(\mathbf{jan})$	def. of A, IS
= $(((\lambda M_{(et)(et)}.\lambda A_{et}.\lambda y_e.M(A)(y) \wedge A(y))(\mathbf{short}))(\mathbf{man}))(\mathbf{jan})$	def. of R_m
= $0 \wedge 1 = 0$	func. application + denotations in M

Intuitively, *Jan* can be a man who is considered to be short in the population of Dutch men, hence $(\mathbf{short}(\lambda x_e.\mathbf{man}(x) \wedge \mathbf{dutch}(x)))\ (\mathbf{jan}) = 1$, but not in the population of all men, hence $(\mathbf{short}\ (\mathbf{man}))(\mathbf{jan}) = 0$. This is a consequence of having *short* denoting a non-intersective modifier: the set denoted by $\mathbf{short}(\lambda x_e.\mathbf{man}(x) \wedge \mathbf{dutch}(x))$ is not necessarily a subset of $\mathbf{short}(\mathbf{man})$.

- Analysis of Pair 2

Let us assume parse trees as follows:

- Text: *Jan* [*is* [*a* [*black* [*Dutch man*]]]]
- Hypothesis: *Jan* [*is* [*a* [*black man*]]]

In analyzing this pair we can show a proof of entailment. Let M be an intended model,

$[\![$ *Jan* [*is* [*a* [*black* [*Dutch man*]]]] $]\!]^M$

= $(\text{IS}(\text{A}((I_m(\mathbf{black}))((I_m(\mathbf{dutch}))(\mathbf{man})))))(\mathbf{jan})$	analysis
= $(((\lambda A_{et}.\lambda B_{et}.\lambda y_e.B(y) \wedge A(y))(\mathbf{black}))(((\lambda A_{et}.\lambda B_{et}.\lambda x_e.B(x) \wedge A(x))\ (\mathbf{dutch}))(\mathbf{man})))(\mathbf{jan})$	def. of A, IS, I_m
= $\mathbf{dutch}(\mathbf{jan}) \wedge (\mathbf{man}(\mathbf{jan}) \wedge \mathbf{black}(\mathbf{jan}))$	func. application
$\leq$ $\mathbf{man}(\mathbf{jan}) \wedge \mathbf{black}(\mathbf{jan})$	def. of $\wedge$
= $(\text{IS}(\text{A}((I_m(\mathbf{black}))(\mathbf{man}))))(\mathbf{jan}) = [\![$ *Jan* [*is* [*a* [*black man*]]] $]\!]^M$	beta reduction + def. of I_m, A, IS + analysis

In this case we rely on the intersectivity of *black*, which in conjunction with the intersectivity of *Dutch* licenses the inference that the set characterized by the et function $[\![$ *black* [*Dutch man*] $]\!]^M$ equals to the set characterized by $[\![$ *Dutch* [*black man*] $]\!]^M$, which is a subset of the set characterized by $[\![$ *black man* $]\!]^M$.

5 Current Annotation Scheme

In the first stages of our attempt to implement the theoretical model described above, we faced a practical problem concerning the binding of expressions in the RTE data to structurally equivalent expressions in the interpreted lexicon: we currently lack an annotation scheme and a user interface that allows annotators to consistently and effectively annotate RTE data. The root of this problem lies in the intricate ways in which the semantic phenomena that we are concerned with are combined with other phenomena or with each other. Simplifying RTE material to an extent that allows binding it to the lexicon as in the above example is often not straightforward. Consider the following example:

Example 4

- T: *Comdex – once among the world's largest trade shows, the launching pad for new computer and software products, and a Las Vegas fixture for 20 years - has been canceled for this year.*
- H: *Las Vegas hosted the Comdex trade show for 20 years.*[9]

Validating the entailment in this pair requires a lexical alignment between an expression in the text and the word *hosted* in the hypothesis. However, there is no expression in the text to establish this alignment. In the text, the noun *Comdex* is in an appositive relation with three conjoined propositions: (i) *once among the world's largest trade shows*; (ii) *the launching pad for new computer and software products*; and (iii) *a Las Vegas fixture for 20 years*. The third element contains a locative restrictive modification in which *Las Vegas* modifies *fixture*. The apposition licenses the inference that *Comdex* is *a Las Vegas fixture* and serves as a prerequisite for the alignment: *Comdex is a Las Vegas fixture* $\Rightarrow$ *Las Vegas hosted Comdex* that simplifies the lexical inference. This alignment is also required for validating the modification by the temporal prepositional phrase *for 20 years* which in the text modifies a noun, *fixture*, and in the hypothesis modifies a verb, *host* - apparently two unrelated lexical items. This example illustrates the difficulty in separating lexical inferences from the semantic relations that underlie the constructions they appear in. In this sense, the manual annotation process that we exemplified in Section 4, in which the stage of *Phenomena Simplification* takes place before the semantic machinery applies, is challenging and requires further investigation with RTE data in order to see what part of the RTE can be annotated using this paradigm, and what elements are needed in order to extend its coverage.

Due to this challenge, and in order to enhance our understanding of the phenomena in the RTE corpora, we adopted a narrower annotation scheme that was carried out on RTE 1-4, named SemAnTE 1.0 - *Semantic Annotation of Textual Entailment*.[10] In this annotation work we focused on valid entailments involving restrictive, intersective and appositive modification that contribute to the recognition of the entailment.[11] In this approach, a construction is annotated if its semantics are required for validating the entailment, but no account is made of the compositional method in which the meaning of the full sentence is obtained. Annotations were marked in 80.65% of the entailments in the RTE 1-4 corpora and reached cross-annotator agreement of 68% on average in four consistency checks. The internal structure of the annotated XML files and a use-case of the annotations for evaluating an entailment component in the BIUTEE recognizer (Stern and Dagan, 2011) are presented in Toledo et al. (2012). See Garoufi (2007) for other relevant work on semantic analysis and annotation of textual entailment done on RTE 2.

5.1 Phenomena Annotated

Our annotations mark inferences by aligning strings in the text and the hypothesis. This is done by pairing each annotation in the text with a corresponding annotation in the hypothesis that marks the output of the inferential process of the phenomenon in question. In the rest of this section we illustrate the phenomena and underline the annotated part in the text with its correspondence in the hypothesis.

[9] Pair 214 from the development set of RTE 1.(Dagan et al., 2006)

[10] The annotated files of SemAnTE are publicly available for download at http://sophia.katrenko.com/CorpusDownload/

[11] Annotators were instructed to construct a full inferential process informally and then to recognize the contribution of the phenomena we aimed to annotate. This method could be applied efficiently only to valid entailments. Invalid entailments marked as *unknown* exhibit an unidentified relation between the text and hypothesis, and pairs marked as *contradictory* rarely center upon the phenomena in question.

5.2 Restrictive modification (RMOD)

- T: *A Cuban$_{Modifier}$ American$_{Modifiee}$ who is accused of espionage pleads innocent.*
- H: *American accused of espionage*.

In this case, *Cuban* modifies *American* and restricts the set of Americans to Cuban Americans. This instance of RMOD validates the inference from *Cuban American* to *American* which is required for establishing the entailment. The intersective nature of the process is not exploited in the actual inference, since the hypothesis does not report that the accused person is Cuban. Thus, only the restrictive property of the modifier *Cuban* is here relevant for the validity of the entailment. More syntactic configurations:

- A verb phrase restricted by a prepositional phrase:
  - T: *The watchdog International Atomic Energy Agency meets in Vienna$_{Modifiee}$ on September 19$_{Modifier}$.*
  - H: *The International Atomic Energy Agency holds a meeting in Vienna*.
- A noun phrase restricted by a prepositional phrase:
  - T: *U.S. officials have been warning for weeks of possible terror attacks$_{Modifiee}$ against U.S. interests$_{Modifier}$.*
  - H: *The United States has warned a number of times of possible terrorist attacks*.

5.3 Intersective Modification (CONJ)

- T: *Nixon was impeached and became the first president ever to resign on August 9th 1974*.
- H: *Nixon was the first president ever to resign*.

This conjunction intersects the two verb phrases *was impeached* and *became the first president ever to resign*. The entailment relies on a subsumption of the full construction to the second conjunct. In addition to canonical conjunctive constructions, CONJ appears also in Restrictive Relative Clauses whereby the relative clause is interpreted intersectively with the noun being modified:

- T: *Iran will soon release eight British servicemen detained along with three vessels*.
- H: *British servicemen detained*.

5.4 Appositive modification (APP)

- Appositive subsumption (left part):
  - T: *Mr. Conway, Iamgold's chief executive officer, said the vote would be close.*
  - H: *Mr. Conway said the vote would be close.*
- Identification of the two parts of the apposition as referring to one another:
  - T: *The incident in Mogadishu, the Somali capital, came as U.S. forces began the final phase of their promised March 31 pullout.*
  - H: *The capital of Somalia is Mogadishu.*

In addition to appositions, APP is annotated in several more syntactic constructions:

- Non-Restrictive Relative Clauses:
  - T: *A senior coalition official in Iraq said the body, which was found by U.S. military police west of Baghdad, appeared to have been thrown from a vehicle*.
  - H: *A body has been found by U. S. military police*.
- Title Constructions:
  - T: *Prime Minister Silvio Berlusconi was elected March 28 with a mandate to reform Italy's business regulations and pull the economy out of recession*.
  - H: *The Prime Minister is Silvio Berlusconi*.

5.5 Marking Annotations

Given a pair from the RTE in which the entailment relation obtains between the text and hypothesis, the task for the annotators is defined as follows:

Table 2: Counters of annotations in RTE 1-4 separated into development and test sets. $A_{\#}$ indicates the number of annotations, $P_{\#}$ indicates the number of entailment pairs containing an annotation and $P_{\%}$ indicates the portion of annotated pairs relative to the total amount of entailment pairs.

(a) RTE 1

	Dev set			Test set		
Ann.	$A_{\#}$	$P_{\#}$	$P_{\%}$	$A_{\#}$	$P_{\#}$	$P_{\%}$
APP	97	87	31	161	134	34
CONJ	90	79	28	126	112	28
RMOD	180	124	44	243	167	42
Any	367	210	74	530	297	74

(b) RTE 2

	Dev set			Test set		
Ann.	$A_{\#}$	$P_{\#}$	$P_{\%}$	$A_{\#}$	$P_{\#}$	$P_{\%}$
APP	179	149	37	155	135	34
CONJ	141	119	30	161	144	36
RMOD	314	205	51	394	236	59
Any	634	318	80	710	350	88

(c) RTE 3

	Dev set			Test set		
Ann.	$A_{\#}$	$P_{\#}$	$P_{\%}$	$A_{\#}$	$P_{\#}$	$P_{\%}$
APP	188	150	38	166	136	34
CONJ	176	138	35	162	134	34
RMOD	300	201	50	307	193	48
Any	664	329	82	635	328	82

(d) RTE 4

	Test set		
Ann.	$A_{\#}$	$P_{\#}$	$P_{\%}$
APP	259	200	40
CONJ	192	164	33
RMOD	429	271	54
Any	880	413	83

1. Read the data, verify the entailment and describe informally why the entailment holds.
2. Annotate all instances of RMOD, APP and CONJ that play a role in the inferential process.

5.6 Annotation Statistics and Consistency

The annotated corpus is based on the scheme described above, applied to the datasets of RTE 1-4 (Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007, 2008). We report annotation statistics in Table 2 and consistency measurements in Table 3. In each consistency check we picked 50-70 entailment pairs that both annotators worked on independently, and compared the phenomena that were annotated.

5.7 Annotation Platform

We used GATE Developer (Cunningham et al., 2011) to annotate the original RTE XML files. The work was performed in two steps using GATE annotation schemes that correspond to RMOD, APP and CONJ: (1) marking the relevant string in the text using one of GATE's schemes (e.g. a scheme of appositive modification), and (2) - marking a string in the hypothesis that corresponds to the output of the inferential process. The annotation in the hypothesis was done using a dedicated *reference_to* scheme.

5.8 Connection to the interpreted lexicon approach

Consider the following pair from RTE 2:
Example 5

- T: *The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991.*
- H: *Shapour Bakhtiar died in 1991.*

Several entailment patterns in this example can be explained by appealing to the semantics of APP, CONJ and RMOD, as follows:

- APP: The appositive modification in *Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991* licenses the inference that *Shapour Bakhtiar and his secretary Sorush Katibeh were found with their throats cut in August 1991.*
- RMOD: The restrictive modification in *August 1991* licenses a subsumption to *1991.*

Table 3: Results of Four Consistency Checks. Each check examined 50-70 annotated pairs from RTE 1-4. In these four checks 66%, 74.11%, 66.67% and 64.66% of the annotations were identical, respectively. On average, 68.03% of the annotations we checked were identical. The rubric *Incorrect Ann.* presents cases of annotations done with an incorrect scheme or with an incorrect scope. *Ambig.-Struct.* are cases of structural or modifier-attachment ambiguity in the text that led to divergent annotations. *Ambig.-Infer.* are cases of divergent annotations stemming from several possible analyses of the inference. *Ambig.-Scheme* refers to instances of divergent annotations due to unclarity or limited specification in the annotation scheme. The last two measures are reported only for the second, third and forth checks.

Measure	RTE 1	RTE 1+2	RTE 3	RTE 4
Data Source(s)	Dev set	Test sets	Dev+Test sets	Test set
Entailment Pairs	50	70	70	70
Total Ann.	93	112	99	133
Identical Ann.	62	83	66	86
Missing Ann.	2	7	7	10
Incorrect Ann.	10	1	2	2
Ambig.-Struct.	9	16	20	15
Ambig.-Infer.	N/A	8	13	12
Ambig.-Scheme	N/A	0	9	7
Consistency (%)	66.67	74.11	66.67	64.66

- CONJ: The conjunction in *Shapour Bakhtiar and his secretary Sorush Katibeh* licenses a subsumption of this expression to *Shapour Bakhtiar*.

By combining these three patterns, we can infer that *Shapour Bakhtiar was found with his throat cut in 1991*. However, additional world knowledge is required to infer that *found with his throat cut* entails *died*. In our current annotation scheme this inference cannot be handled since lexical alignment of unmodeled phenomena is not supported . This motivates a more robust approach as proposed in Section 4.

6 Conclusions

The goal of this research is to establish a model-theoretic benchmark explaining entailment data. We have presented a model that utilizes standard semantic principles and illustrated the way it accounts for textual entailment from the RTE corpora. The model centers upon an interpreted lexicon that comprises words and operators. These elements are used to represent a fragment of English to which premises and hypotheses may be bound.

We focus on the annotation of semantic phenomena which are predominant in the RTE corpora and can be annotated with high consistency, but which may have several syntactic expressions and therefore allow us to generalize regarding abstract entailment patterns. Non-modeled phenomena that exist in the data are simplified in a preparatory step but cases in which such phenomena are deeply intertwined with the semantic phenomena that we model pose a challenge for the formalization of an annotation scheme.

At a first stage, we carried out a restricted annotation scheme by which instances of restrictive, intersective, and appositive modification are marked in entailment pairs with no account for the full inferential process between the premise and the hypothesis. These phenomena were found in 80.65% of the entailments in RTE 1-4 and were marked with cross-annotator agreement of 68% on average.

We are currently investigating different directions in the formulation of an extensive annotation scheme coincident with the model we described and are aiming to develop a corresponding annotation platform. This platform would allow annotators to bind constructions manifesting supported semantic phenomena to representations in the interpreted lexicon as well as to simplify lexical/syntactic phenomena of the kind illustrated in Examples 2 and 4 by textual alignment. In the next stages of this project, we plan to use an external theorem prover to automatically validate the entailment relation (or lack thereof).

References

Bar Haim, R., I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor (2006). The second pascal recognising textual entailment challenge. In *In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Bentivogli, L., I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini (2009). The fifth pascal recognizing textual entailment challenge. *Proceedings of TAC 9*, 14–24.

Bos, J., S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier (2004). Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 12–40.

Bos, J. and K. Markert (2005). Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 628–635.

Cooper, R., D. Crouch, J. Van Eijck, C. Fox, J. Van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, S. Pulman, T. Briscoe, H. Maier, and K. Konrad (1996). *Using the Framework*. The Fracas Consortium.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters (2011). *Text Processing with GATE (Version 6)*.

Dagan, I., O. Glickman, and B. Magnini (2006). The pascal recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190.

Garoufi, K. (2007). Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. Master's thesis, Saarland University.

Giampiccolo, D., H. T. Dang, B. Magnini, I. Dagan, and E. Cabrio (2008). The fourth pascal recognising textual entailment challenge. In *In TAC 2008 Proceedings*.

Giampiccolo, D., B. Magnini, I. Dagan, and B. Dolan (2007). The third pascal recognizing textual entailment challenge. In *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, Stroudsburg, PA, USA, pp. 1–9. Association for Computational Linguistics.

Kamp, H. and U. Reyle (1993). *From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and discourse representation theory*, Volume 42. Kluwer Academic Dordrecht,, The Netherlands.

MacCartney, B. and C. D. Manning (2007). Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 193–200.

Pratt-Hartmann, I. and L. S. Moss (2009). Logics for the relational syllogistic. *Review of Symbolic Logic*, 647–683.

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

Toledo, A., S. Katrenko, S. Alexandropoulou, H. Klockmann, A. Stern, I. Dagan, and Y. Winter (2012). Semantic annotation for textual entailment recognition. In *Proceedings of the Eleventh Mexican International Conference on Artificial Intelligence (MICAI)*.

Valencia, V. S. (1991). *Studies on natural logic and categorial grammar*. Ph. D. thesis, University of Amsterdam.