# Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference

**Luisa Bentivogli[1], Elena Cabrio[1], Ido Dagan[2],**

**Danilo Giampiccolo[3], Medea Lo Leggio[3], Bernardo Magnini[1]**

[1]FBK- irst, Trento, Italy
[2] Bar-Ilan University, Ramat Gan, Israel
[3]CELCT, Trento, Italy
E-mail: {bentivo,cabrio,magnini}@fbk.eu; dagan@cs.biu.ac.il;
{giampiccolo,loleggio}@celct.it

**Abstract**

This paper proposes a methodology for the creation of specialized data sets for Textual Entailment, made of monothematic Text-Hypothesis pairs (i.e. pairs in which only one linguistic phenomenon relevant to the entailment relation is highlighted and isolated). The annotation procedure assumes that humans have knowledge about the linguistic phenomena relevant to inference, and a classification of such phenomena both into fine grained and macro categories is suggested. We experimented with the proposed methodology over a sample of pairs taken from the RTE-5 data set, and investigated critical issues arising when entailment, contradiction or unknown pairs are considered. The result is a new resource, which can be profitably used both to advance the comprehension of the linguistic phenomena relevant to entailment judgments and to make a first step towards the creation of large-scale specialized data sets.

## 1. Introduction

Recognizing Textual Entailment (RTE) consists of developing a system that, given two text fragments (a text T and a hypothesis H), can determine whether the meaning of one text snippet can be inferred from the other (Dagan *et al.*, 2005). To test the progress of TE systems in a comparable setting, the participants to the RTE Campaign are provided with data sets composed of T-H pairs involving various levels of entailment reasoning (e.g. le11xical, syntactic) and are required to produce a correct judgment on the given pairs. Two kinds of judgments are allowed: two-way (yes or no entailment) or three-way judgment (entailment, contradiction, unknown). To perform the latter, in case there is no entailment between T and H systems must be able to distinguish whether the truth of H is contradicted by T, or remains unknown on the basis of the information contained in T. According to the task proposed, the RTE-4 and RTE-5 data sets are annotated for a 3-way decision: "entailment" (50% of the pairs), "unknown" (35%) and "contradiction" (15%), resulting in 50% positive examples and 50% negative examples.

To correctly judge each single pair inside the RTE data sets, systems are expected to cope both with the different linguistic phenomena involved in TE, and with the complex ways in which they interact.

One of the major issues raised by the TE community is that while system developers create new modules, algorithms and resources to address specific inference types, it is difficult to measure a substantial impact when such modules are evaluated on the RTE data sets because of (i) the sparseness (i.e. low frequency) of the single phenomena, and (ii) the impossibility to isolate each phenomenon, and to evaluate each module independently from the others. A similar expectation emerges from several studies in the literature, where the interest in the development of systems and resources to deal with the different linguistic levels involved in TE comes to light.

According to such considerations, this paper describes a methodology for the creation of specialized TE data sets made of monothematic T-H pairs, i.e. pairs in which a certain phenomenon relevant to the entailment relation is highlighted and isolated. The proposed methodology starts from an existing RTE pair and defines the following steps: (i) identify the phenomena present in the original RTE pair; (ii) apply an annotation procedure to isolate each phenomenon and create the related monothematic pair; finally, (iii) group together all the monothematic T-H pairs relative to the same phenomenon, hence creating specialized data sets.

The expected benefits of specialized data sets for TE derive from the intuition that investigating the linguistic phenomena separately, i.e. decomposing the complexity of the TE problem, would yield an improvement in the development of specific strategies to cope with them. In fact, being able to detect entailment basing on linguistic foundations should strengthen the systems, making the overall performances less data set dependent.

We carried out a feasibility study applying the devised methodology to a sample of 90 pairs extracted from the RTE-5 data set (Bentivogli *et al.*, 2009) and we addressed a number of critical issues, including:

- whether it is possible to clearly identify and isolate the linguistic phenomena underlying the entailment relation;
- how specific the categorization of phenomena should be;
- how easy/difficult it is to create balanced data sets of monothematic T-H pairs with respect to the distribution of positive and negative examples, so that these data sets might be used for training and testing.

The result of the feasibility study is a "pilot" resource

(freely available at the Textual Entailment Resource Pool website[1]), which can be profitably used both to advance in the comprehension of the linguist phenomena involved in the entailment judgments, and to make a first step toward the creation of large-scale specialized data sets.

The paper is structured as follows. Section 2 reports on previous work related to specialized data sets for textual entailment, which constitutes the starting point of this paper. Section 3 describes the annotation methodology for the creation of the specialized data sets, highlighting the linguistic phenomena which are detected and isolated, and describing in detail the procedure devised to create the monothematic pairs. In Section 4 some examples of the application of the methodology are presented, while in Section 5 a feasibility study carried out on a sample of the RTE-5 data set is described and the resulting data are given. In Section 6 a number of issues that arise while trying to create a balanced data set are presented, and Section 7 concludes the paper drawing some remarks and discussing on the feasibility of the proposed approach for the creation of large scale data sets.

## 2. Related work

The interest of the research community in producing specific resources to deal with linguistic phenomena underlying the entailment relation is proven by a number of different works in the field. Several studies in the literature (e.g. Vanderwende *et al.*, 2005; Clark *et al.*, 2007) point out that the lexical, syntactic and world knowledge levels can be analyzed and exploited in order to fully identify and recognize the entailment between T and H.

In (Garoufi, 2007), a scheme for manual annotation of textual entailment data sets called ARTE is proposed, with the aim of highlighting a wide variety of entailment phenomena in the data. ARTE views the entailment task in relation to three levels, i.e. Alignment, Context and Coreference, according to which 23 different features for positive entailment annotation are extracted. Each level is explored in depth for the positive entailment cases, while for the negative pairs a more basic scheme is conceived. The ARTE scheme has been applied to the complete positive entailment RTE-2 Test Set (400 pairs), and to a random 25% portion of the negative entailment Test Set.

As part of their work on detecting contradictions in text, Stanford NLP Group created a corpus where contradictions arise from negation by adding negative markers to the RTE-2 test data[2]. In this corpus, a single phenomenon is investigated, namely negation (de Marneffe *et al.*, 2008). Kirk (2009) describes his ongoing work of building an inference corpus for spatial inference about motion, while (Akhmatova & Dras, 2009) shows how current approaches on hypernymy acquisition can be adapted to improve entailment classification accuracy.

A step further is carried out in (Magnini & Cabrio, 2009) which defines a general method for the combination of specialized entailment engines (EEs), each of which able to deal with a certain aspect of language variability. In order to train the implemented specialized EEs, they define and use specialized Text-Hypothesis data sets for a certain linguistic phenomenon *i* (notated as $[T,H]_i$), made of monothematic Text-Hypothesis pairs which, according to human judgments, can be resolved by means of a single linguistic phenomenon *i*. As a first attempt to prove the feasibility of their approach, two small (i.e. 50 pairs each) monothematic data sets are created, one for negation ($[T,H]_{neg}$) and one for lexical similarity ($[T,H]_{lex}$), partially derived from T-H pairs included in the RTE data. The data sets are balanced between positive and negative entailment and are aligned with respect to their Ts, while the Hs are built removing all the linguistic phenomena but the one under consideration.

## 3. Methodology

In this work we extend and refine the methodology defined in (Magnini & Cabrio, 2009), with the aim of applying it systematically to the RTE-5 data set. The idea is to create monothematic pairs on the basis of the phenomena which are actually present in the RTE-5 T-H pairs. One of the advantages of applying the methodology to the RTE data consists of the fact that the actual distribution of the linguistic phenomena involved in the entailment relation emerges. In the following, we propose a classification of the linguistic phenomena (Section 3.1) and describe the entailment rules according to which we create the monothematic pairs (Section 3.2); we then explain the procedure followed for their creation (Section 3.3).

### 3.1. Classification of linguistic phenomena

We grouped linguistic phenomena using both fine-grained categories and broader categories. Grouping specific phenomena into macro categories will allow us to create specialized data sets containing enough pairs to train and test TE systems. Macro categories are defined referring to widely accepted linguistic categories in the literature (e.g. Garoufi, 2007) and to the inference types typically addressed in RTE systems: *lexical, syntactic, lexical-syntactic, discourse and reasoning*.

Each macro category includes fine-grained phenomena, which are listed below. This list is not exhaustive and reflects the phenomena we detected in the sample of RTE-5 pairs we analyzed.

- *lexical*: identity, format, acronymy, demonymy, synonymy, semantic opposition, hyperonymy, geographical knowledge;
- *lexical-syntactic*: transparent heads, nominalization/verbalization, causative, paraphrase;
- *syntactic*: negation, modifier, argument realization, apposition, list, coordination, active/passive alternation;
- *discourse*: coreference, apposition, zero anaphora, ellipsis, statements;
- *reasoning*: apposition, modifiers, genitive, relative clause, elliptic expressions, meronymy, metonymy, membership/representativeness, reasoning on

[1] http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

[2] http://www-nlp.stanford.edu/projects/contradiction/

quantities, temporal and spatial reasoning, all the general inferences using background knowledge.

Some phenomena (e.g. apposition) can be classified in more than one macro category, according to their specific occurrence in the text. For instance (pair 8 in RTE-5):

T: *The government of Niger and Tuareg rebels of the Movement of Niger People for Justice (MNJ) have agreed to end hostilities [...].*
H: *MNJ is a group of rebels.*

the apposition is considered as syntactic, while in pair 28:

T: *Ernesto, now a tropical storm, made landfall along the coastline of the state of North Carolina [...].*
H: *Ernesto is the name given to a tropical storm.*

the apposition is classified into the category reasoning.

It is worthwhile to note that since world knowledge is an omni-pervasive phenomenon, it has not been categorized separately; this aspect can be considered as an open issue that needs further investigation. Details on the distribution of each phenomenon in the sample are given in Section 5.

### 3.2. Entailment rules

We assume that humans have knowledge about the linguistic phenomena relevant to textual entailment, and that such knowledge can be expressed through entailment rules (Szpektor et al., 2007).

An entailment rule is either a directional or bidirectional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees, according to the granularity of the phenomenon they formalize). The left-hand side of the pattern (LHS) entails the rights-hand side (RHS) of the same pattern under the same variable instantiation. In addition, a rule may be defined by a set of constraints, representing variable typing (e.g. PoS, Named Entity type) and relations between variables, which have to be satisfied for the rule to be correctly applied. For instance, the entailment rule for demonyms can be expressed as:

Pattern: *X Y* ↔ *Y* (IS) FROM *Z*
Constraint: *DEMONYM(X,Z)*
*TYPE(X)=ADJ_NATIONALITY; TYPE(Z)=GEO*

meaning that *x y* entails *y is from z* if there is a ENTAILMENT relation of demonymy between *x* and *y*, *x* is an adjective expressing a nationality and *z* is a geographical entity (e.g. *A team of European astronomers* ↔ *A team of astronomers from Europe*, pair 205 RTE-5).

The entailment rules for a certain phenomenon aim to be as general as possible, but for the cases in which the semantics of the words is essential (e.g. general inference), text snippets extracted from the data are used. Different rules can be needed in order to formalize the variants in which the same phenomenon occurs in the pairs. For example, both the following entailment rules formalize the phenomenon of apposition (syntax):

a) Pattern: *X Y* ↔ *Y X*
Constraint: *APPOSITION(Y,X)*

b) Pattern: *X, Y* ↔ *Y IS X*
Constraint: *APPOSITION(Y,X)*

### 3.3. Procedure for the creation of monothematic pairs

The procedure consists of a number of steps carried out manually. We start from a [*T,H*] pair taken from one of the RTE data sets and we decompose [*T,H*] in a number of monothematic pairs [$T,H_i$], where *T* is the original Text and $H_i$ are Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in [*T,H*]. The procedure is schematized in the following steps:

***1.*** Individuate the linguistic phenomena which contribute to the entailment in [*T,H*]

***2.*** For each phenomenon *i*:

a) individuate a general entailment rule $r_i$ for the phenomenon *i,* and instantiate the rule using the portion of T which expresses *i* as the LHS of the rule, and information from H on *i* as the RHS of the rule.

b) substitute the portion of T that matches the LHS of $r_i$ with the RHS of $r_{i.}$.

c) consider the result of the previous step as $H_i$, and compose the monothematic pair [$T,H_i$]. Mark the pair with phenomenon *i*.

***3.*** Assign an entailment judgment to each monothematic pair.

After applying this procedure to the original pairs, all the monothematic *[T-Hi]* pairs relative to the same phenomenon *i* should be grouped together in a data set specialized for phenomenon *i*.

## 4. Application of the methodology

In this section, we show examples of the application of the procedure to RTE pairs.

### 4.1. Entailment pairs

Table 1 shows the decomposition of an original entailment pair (pair 408 in RTE-5) into monothematic pairs. At step 1 of the methodology, the linguistic phenomena (i.e. apposition, synonymy, verbalization and argument realization) are considered relevant to the entailment between T and H. In the following, we apply step by step the procedure to the phenomenon we define as argument realization. At step 2a the general rule:

Pattern: *X Y* ↔ *Y* IN *X*
Constraint: *TYPE(X) = TEMPORAL_EXPRESSION*

is instantiated (*2007 Nobel Prize in Literature* ↔ *Nobel Prize in Literature in 2007*), while at step 2b the substitution in T is carried out *([...] Doris Lessing, recipient of the Nobel Prize (in Literature) in 2007 [...]*[3]).

[3] The symbol […] is used as a placeholder of the missing parts.

| Pair 408 | | | Text snippet | Rule | Phenomena | Judg. |
|---|---|---|---|---|---|---|
| T | | | British writer Doris Lessing, recipient of the 2007 Nobel Prize in Literature, has said in an interview that the terrorist attack on September 11 "wasn't that terrible" when compared to attacks the Irish Republican Army (IRA) made on Britain. […] | | | |
| H | | | Doris Lessing won the Nobel Prize in Literature in 2007. | | **syntax:argument realization, syntax:apposition, lexical:verbalization, lexical:synonymy** | E |
| | H1 | | British writer Doris Lessig, recipient of the Nobel Prize in Literature **in 2007**, has said that the terrorist attack on September 11 "wasn't that terrible". […] | **x y ↔ y in x** ***Type(x)=temporal_expression*** | **syntax:argument realization** | E |
| | H2 | | **British writer Doris Lessing is the recipient** of the 2007 Nobel Prize in Literature. | **x, y → y is x** ***apposition(x, y)*** | **syntax:apposition** | E |
| | H3 →**T'** | | British writer Doris Lessing **received** the 2007 Nobel Prize in Literature. | **x → y** ***Type(x)=N*** ***Type (y)=V*** ***verbalization_of (y,x)*** | **lexical:verbalization** | E |
| | | H4 | British writer Doris Lessing **won** the 2007 Nobel Prize in Literature. | **x ↔ y** ***synonym(x,y)*** | **lexical:synonymy** | E |

Table 1: Example of the application of the methodology to an entailment pair.

| Pair 125 | | | Text snippet | Rule | Phenomena | Judg. |
|---|---|---|---|---|---|---|
| T | | | **Mexico's new president, Felipe Calderon,** seems to be doing all the right things in cracking down on Mexico's drug traffickers. […] | | | |
| H | | | Felipe Calderon is the outgoing President of Mexico. | | **lexical:semantic opposition, syntactic:argument realization, syntactic:apposition** | C |
| | H1 | | Mexico's **outgoing** president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. […] | **x ←/→ y** ***antonym(x,y)*** | **lexical :semantic opposition** | C |
| | H2 | | **The new president of Mexico**, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. . […] | **x's y → y of x** | **syntactic:argument realization** | E |
| | H3 | | **Felipe Calderon is** Mexico's **new president**. | **x, y → y is x** ***apposition(y,x)*** | **syntactic:apposition** | E |

Table 2: Example of the application of the methodology to a contradiction pair.

At step 2c the monothematic pair T-$H_1$ is composed and marked as "argument realization" (macro-category "syntactic"). Finally, at step 3, this pair is judged as "entailment". Step 2 (a, b, c) is then repeated for all the phenomena individuated in that pair at step 1.

It can be the case that several phenomena are collapsed on the same token. For instance, in Table 1 a chain of two phenomena should be solved to match *recipient of* with *won*. In such cases, in order to create a monothematic H for each phenomenon, the methodology is applied recursively. It means that after applying it once to the first phenomenon of the chain (therefore creating the pair $T-H_i$), it is applied again on $H_i$ (that becomes $T'$) to solve the second phenomenon of the chain (creating the pair $T'-H_j$); more specifically, in Table 1 the methodology is first applied on T for the verbalization ($T-H_3$) and then, it is recursively applied on $H_3$ (that becomes $T'$) to solve the synonymy ($T'-H_4$).

## 4.2. Contradiction pairs

Table 2 shows the decomposition of an original contradiction pair (pair 125 in RTE-5) into monothematic pairs. At step 1 both the phenomena that preserve the entailment and the phenomena that break the entailment rules causing a contradiction in the pair should be detected.

In the example reported in Table 2, the phenomena that should be solved in order to correctly judge the pair are: argument realization, apposition and semantic opposition. While the monothematic pairs created basing on the first two phenomena preserve the entailment, the semantic opposition generates a contradiction. In the following, we apply step by step the procedure to the phenomenon of semantic opposition. At step 2a the general rule:

Pattern: $X \leftarrow/\rightarrow Y$
Constraint: *SEMANTIC_OPPOSITION(Y,X)*

is instantiated (*new*←/→*outgoing*), and at step 2b the substitution in T is carried out (*Mexico's outgoing president, Felipe Calderon [...]*). At step 2c a *negative* monothematic pair T-$H_1$ is composed and marked as semantic opposition (macro-category "lexical"), and the pair is judged as "contradiction".

We noticed that negative monothematic T-*H* pairs (i.e. both contradiction and unknown) may originate either from the application of contradiction rules (e.g. semantic opposition or negation, as in pair $T\text{-}H_1$, in Table 2) or as a wrong instantiation of a positive entailment rule. For instance, the positive rule for active/passive alternation:

Pattern: $X\ Y\ Z \leftrightarrow Z\ \mathrm{W}\ X$
Constraint: *SAME_STEM(X,W)*
*TYPE(X)=V_ACT; TYPE(W)=V_PASS*

when wrongly instantiated, as in *Russell Dunham killed nine German soldiers −/→ Russell Dunham was killed by nine German soldiers* ($X\ Y\ Z \leftrightarrow Z\ \mathrm{W}\ X$), generates a negative monothematic pair.

## 4.3. Unknown pairs

Table 3 shows the decomposition of an original unknown pair (pair 82 in RTE-5) into monothematic pairs. At step 1 all the relevant phenomena are detected: coreference, background knowledge, and modifier.

While the first two preserve the entailment relation, the monothematic pair resulting from the third phenomenon is judged as unknown. In the following, we apply step by step the procedure to the phenomenon of modifier.

| **Pair 82:** | | | **Text snippet** | **Rule** | **Phenomena** | **Judg.** |
|---|---|---|---|---|---|---|
| T | | | Currently, there is no specific treatment available against **dengue fever, which is the most widespread tropical disease after malaria**. Sanofi Pasteur is collaborating with the Communicable Disease Center in Singapore and the Pasteur Institute in Vietnam to conduct these clinical studies in children and adults. "Controlling **the mosquitoes that transmit dengue** is necessary but not sufficient to fight against the disease. […]" | | | |
| H | | | Malaria is the most widespread disease transmitted by mosquitoes. | | **discourse:coreference, reasoning:background knowledge, syntax:modifier** | U |
| | H1 →**T'** | | **Dengue fever** is the most widespread tropical disease after malaria. | **x ↔ y** ***coreference(x,y)*** | **discourse:coreference** | E |
| | | H2 | **Malaria is the most widespread** tropical disease. | **x is the second after y → y is the first** | **reasoning:background knowledge** | E |
| | | H3 | Dengue fever is the most widespread tropical disease **trasmitted by mosquitos** after malaria. | **x −?→ x y** ***modifier(y,x)*** | **syntax:modifier (restr. relative clause)** | U |

Table 3: Example of the application of the methodology to an unknown pair.

In detail, at step 2a the generic rule:

Pattern: $X -?\rightarrow X\ Y$
Constraint: $MODIF(Y,X)$

is instantiated (*disease−?→ disease transmitted by mosquitoes*), and at step 2b the substitution in T is carried out. At step 2c the monothematic pair T'-$H_3$ is composed and marked as "modifier" (restrictive relative clause, macro-category "lexical"), and the pair is judged as "unknown".

## 5. Feasibility Study on RTE-5 Data

In order to assess the feasibility of the specialized data sets, we applied our methodology to a sample of 90 T-H pairs randomly extracted from the RTE-5 data set. In particular, the sample pairs are equally taken from "entailment", "contradiction" and "unknown" examples.

The whole RTE-5 sample has been annotated by two annotators with skills in linguistics and inter-annotator agreement has been calculated. A first measure of "complete" agreement was considered, counting when judges agree on all phenomena present in a given original T-H pair. The complete agreement on the full sample amounts to 64.4% (58 up to 90 pairs).

In order to account for partial agreement on the set of phenomena present in the T-H-pairs, we used the Dice coefficient (Dice, 1945)[4]. The Dice coefficient is computed as follows:

Dice=2C/(A+B)

where C is the number of common phenomena chosen by the annotators, while A and B are respectively the number of phenomena detected by the first and the second annotator. Inter-annotator agreement on the whole sample amounts to 0.78. Overall, we consider this value high enough to demonstrate the stability of the (micro and macro) phenomena categories, thus validating their classification model.

Table 4 shows inter-annotator agreement rates grouped according to the type of the original pairs, i.e. "entailment", "contradiction" and "unknown" pairs.

| | **Complete** | **Partial (Dice)** |
|---|---|---|
| ENTAILMENT | 60% | 0.86 |
| CONTRADICTION | 57% | 0.75 |
| UNKNOWN | 76% | 0.68 |

Table 4. Agreement measures per entailment type

The highest percentage of "complete" agreement is obtained on "unknown" pairs. This is due to the fact that since the H in "unknown" pairs typically contains information which is not present in (or inferable from) T, for 19 pairs out of 30 both the annotators agreed that no linguistic phenomena relating T to H could be detected.

With respect to the Dice coefficient, the highest inter-annotator agreement can be seen for the "entailment" pairs, whereas the agreement rates are lower for "contradiction" and "unknown" pairs. This is due to the fact that for the "entailment" pairs, all the single phenomena are directly involved in the entailment relation, making their detection straightforward. On the contrary (cfr. Sections 4.2 and 4.3), in the original "contradiction" and "unknown" pairs not only the phenomena directly involved in the contradiction/unknown relation are to be detected, but also those preserving the entailment, which do not play a direct role on the relation under consideration (contradiction/unknown) and are thus more difficult to be identified.

The distribution of the phenomena present in the original RTE-5 pairs, as resulting after a reconciliation phase carried out by the annotators, is shown in Table 5.

The total number of occurrences of each specific phenomenon is given (Column *TOT*), corresponding to the number of monothematic pairs created for that phenomenon. The number of monothematic pairs is then broken down into positive examples - i.e. "entailment" monothematic pairs (Column *E*) - and negative examples – i.e. "contradiction" and "unknown" monothematic pairs (Columns *C* and *U*, respectively).

A number of remarks can be made on the data presented in Table 5. Both macro categories and fine-grained phenomena are well represented but show a different absolute frequency: some have a high number of occurrences, whereas some others occur very rarely. In particular, as already pointed out in (Garoufi, 2007), also our study confirms that the phenomena belonging to the category *reasoning* are the most frequent, meaning that a significant part of the data involves deeper inferences.

As for the distribution among E/C/U monothematic pairs, we can see that some phenomena appear more frequently - or only - among the positive examples (e.g. apposition or coreference) and others among the negative ones (e.g. quantitative reasoning). In general, the total number of positive examples is much higher than that of the negative ones and, for some macro-categories (e.g. lexical-syntactic) no negative examples are found. Also from a qualitative standpoint, the variability of phenomena in negative examples is reduced with respect to the positive pairs.

Overall, the feasibility study showed that the decomposition methodology proposed in this paper can be applied on RTE-5 data. The task demonstrated to be feasible under a number of aspects. As for the quality of the monothematic pairs, the high inter-annotator agreement rate obtained shows that the methodology is stable enough to be applied on a large scale. With respect to the human effort required, during the feasibility study an average of

[4]The Dice coefficient is a typical measure used to compare sets in IR and is also used to calculate inter-annotator agreement in a number of tasks where an assessor is allowed to select a set of labels to apply to each observation. In fact, in these cases, and in ours as well, measures such as the widely used K are not good to calculate agreement. This is because K only offers a dichotomous distinction between agreement and disagreement, whereas what is needed is a coefficient that also allows for partial disagreement between judgments.

four original RTE-5 pairs per hour have been decomposed. This means that, provided that the task be carried out by annotators with a curriculum in linguistics, around two and a half person months are required to apply the decomposition methodology to the whole RTE-5 data set, which is composed of 1,200 T-H pairs.

| Phenomena | Monothematic Pairs | | | |
|---|---|---|---|---|
| | TOT | E | C | U |
| **Lexical:** | **32** | **22** | **8** | **2** |
| Identity/mismatch | 4 | 1 | 3 | 0 |
| Format | 2 | 2 | 0 | 0 |
| Acronymy | 3 | 3 | 0 | 0 |
| Demonymy | 1 | 1 | 0 | 0 |
| Synonymy | 11 | 11 | 0 | 0 |
| Semantic opposition | 3 | 0 | 3 | 0 |
| Hypernymy | 5 | 3 | 0 | 2 |
| Geographical knowledge | 3 | 1 | 2 | 0 |
| **Lexical-syntactic** | **18** | **18** | **0** | **0** |
| Transparent head | 3 | 3 | 0 | 0 |
| Nominalization/verbalization | 9 | 9 | 0 | 0 |
| Causative | 1 | 1 | 0 | 0 |
| Paraphrase | 5 | 5 | 0 | 0 |
| **Syntactic** | **44** | **30** | **10** | **4** |
| Negation | 1 | 0 | 1 | 0 |
| Modifier | 8 | 4 | 1 | 3 |
| Argument realization | 6 | 6 | 0 | 0 |
| Apposition | 17 | 11 | 6 | 0 |
| List | 1 | 1 | 0 | 0 |
| Coordination | 5 | 4 | 0 | 1 |
| Active/passive alternation | 6 | 4 | 2 | 0 |
| **Discourse** | **44** | **43** | **0** | **1** |
| Coreference | 24 | 23 | 0 | 1 |
| Apposition | 3 | 3 | 0 | 0 |
| Anaphora zero | 12 | 12 | 0 | 0 |
| Ellipsis | 4 | 4 | 0 | 0 |
| Statements | 1 | 1 | 0 | 0 |
| **Reasoning** | **67** | **45** | **17** | **6** |
| Apposition | 3 | 2 | 1 | 0 |
| Modifier | 3 | 3 | 0 | 0 |
| Genitive | 1 | 2 | 0 | 0 |
| relative clause | 1 | 1 | 0 | 0 |
| elliptic expression | 1 | 1 | 0 | 0 |
| Meronymy | 4 | 3 | 1 | 0 |
| Metonymy | 3 | 3 | 0 | 0 |
| membership/representative | 2 | 2 | 0 | 0 |
| Quantity | 6 | 0 | 5 | 1 |
| Temporal | 2 | 1 | 0 | 1 |
| Spatial | 1 | 1 | 0 | 0 |
| common background/ general inferences | 40 | 26 | 10 | 4 |
| **TOTAL (# monothematic pairs)** | **206** | **158** | **35** | **13** |

Table 5: distribution of phenomena in T-H pairs

## 6. Creating Specialized Data sets

After applying the procedure described in Section 3.3 to the original 90 pairs of our sample, all the monothematic *[T-Hi]* pairs relative to the same phenomenon *i* can be grouped together, resulting in several data sets specialized for phenomenon *i*. For instance, we can create a specialized data set for Reasoning phenomena, which would include 67 monothematic pairs, out of which 45 are positive, 17 are contradiction and 6 are unknown (see Table 5).

As introduced before, due to the natural distribution of phenomena in RTE data, we found out that applying the decomposition methodology we generate a higher number of monothematic positive pairs (76.7%) than negative ones (23.3%, divided into 17% "contradiction" and 6.3% "unknown", as shown in Table 5).

We analyzed separately the three subsets composing the RTE-5 sample, (i.e. 30 "entailment" pairs, 30 contradiction pairs, and 30 "unknown") in order to verify the productivity of each subset with respect to the monothematic pairs created from them. Table 6 shows the absolute distribution of the monothematic pairs among the three RTE-5 classes.

| Original RTE-5 pairs | | Phenomena / monothematic pairs | | | |
|---|---|---|---|---|---|
| | | E | C | U | Total |
| | E (30) | 91 | -- | -- | 91/30 |
| | C (30) | 44 | 35 | -- | 79/30 |
| | U (30) | 23 | -- | 13 | 36/11 |

Table 6: distribution of the monothematic pairs with respect to original E/C/U pairs

When the methodology is applied to RTE-5 "entailment" examples, averagely 3.03 all positive monothematic pairs are derived.

When the methodology is applied to RTE-5 "contradiction" examples, we can create an average of 2.64 monothematic pairs, among which 1.47 are entailment pairs and 1.17 are contradiction pairs. This means that the methodology is productive for both positive and negative examples.

As introduced before, in 19 out of 30 "unknown" examples no monothematic pairs can be created, due to the lack of specific phenomena relating T and H (typically the H contains information which is neither present in T or inferable from it). For the 11 pairs that have been decomposed into monothematic pairs, we created an average of 3.27 monothematic pairs, among which 2.09 are entailment pairs and 1.18 are unknown pairs.

This analysis shows that the only source of negative monothematic pairs are the RTE-5 "contradiction" pairs, which actually correspond to 15% of the RTE-5 data set.

As regards the issue of balancing each single specialized data set with respect to positive and negative examples (i.e. finding a balanced number of positive and negative examples for each single phenomenon) we saw in Section 5 that some phenomena appear more frequently – when not only - among the positive examples (e.g. apposition or coreference) while others appear more among the negative ones (e.g. quantitative reasoning). It happens that not only for specific phenomena but also for

entire macro categories (e.g. lexical-syntactic) negative examples cannot be found.

Although the specialized data sets derived from the decomposition procedure might be useful for interesting corpus analysis investigations, current systems based on machine learning approaches would benefit from data sets with a more balanced proportion of negative examples. To cope with this problem, we devised a tentative solution, which consists of taking a positive example for a given phenomenon and synthetically creating a corresponding negative example by modifying the entailment rule.

Starting from the observation of original "contradiction" and "unknown" pairs described in Section 4.2 and 4.3., we spotted out some possible operations to invalidate the rule which preserves the entailment in positive examples:

(i) invert a directional rule

Ex (pair 187 – REASONING:MODIFIER):
T: *[...] Islands are mostly made up of mangrove trees.*
H1-pos: *Mangroves are a kind of tree.*
H1-neg: *Trees are a kind of mangrove*.

(ii) wrongly instantiate a rule

Ex (pair 408, cfr Table 1 – LEXICAL:VERBALIZATION)
T: *[...]Doris Lessing, recipient of the 2007 Nobel Prize [...]*
H3-pos: *Doris Lessing received the 2007 Nobel Prize*
H3-neg: *Doris Lessing receipted the 2007 Nobel Prize*

In this example the verbalization rule is wrongly instantiated by using a verb with the same stem of the verb "receive" but with another meaning.

(iii) where possible, substitute the rule with another rule related to an opposite phenomenon.

Ex (pair 408, cfr. Table 1 – LEXICAL:SYNONYMY)
T': *[...] Doris Lessing received the 2007 Nobel Prize [...]*
H4-pos: *Doris Lessing* ***won*** *the 2007 Nobel Prize*
H4-neg: *Doris Lessing refused the 2007 Nobel Prize*

This operation exploits the natural opposition of some phenomena (e.g. identity vs. negation; synonymy vs. oppositeness). In the example, the verb "win", which is synonym of "receive" is substituted with the verb "refuse", which is semantically opposed to "receive".

Two annotators carried out a feasibility study on the RTE-5 sample and found out that it was a difficult and time-consuming task leading to low inter-annotator agreement. For this reason, we suggest that alternative strategies for the generation of negative monothematic pairs be further discussed. How to collect more negative examples is still an open issue for which we advocate attention from the research community.

## 7. Conclusions

This paper presents a methodology for the creation of specialized TE data sets, made of monothematic T-H pairs in which a certain phenomenon underlying the entailment relation is highlighted and isolated. We carried out a pilot study applying such methodology to a sample of 90 pairs extracted from the RTE-5 data set and we demonstrated the feasibility of the task, both in terms of quality of the new pairs created and of time and effort required.

An important outcome of the methodology proposed is that we provide the annotation of previous RTE data with the linguistic phenomena underlying the entailment/contradiction relations in the pairs (both with fine grained and macro categories), highlighting their actual distribution in the data, and allowing evaluations of the TE systems on specific phenomena both when isolated and when interacting with the others. Unlike previous work of analysis on RTE data, the result of our study is a new resource that can be used for training TE systems on specific linguistic phenomena relevant to inference. Finally, in order to face the emerged problem of finding enough negative examples to be included in the specialized data sets, a first attempt to define a procedure for the creation of negative monothematic pairs has been presented.

## 8. References


Akhmatova E., Dras M. (2009). Using Hypernymy Acquisition to Tackle (Part of) Textual Entailment. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009)*, Singapore.

Bentivogli, L., Dagan, I., Dang H.T., Giampiccolo, D., Magnini, B. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the TAC Workshop*, Gaithersburg, MD, USA.

Clark, P., Harrison, P., Thompson, J., Murray, W., Hobbs, J., and Fellbaum, C. (2007). On the Role of Lexical and World Knowledge in RTE3. In *Proceedings of ACL-PASCAL Workshop on TE and Paraphrasing.* Prague, Czech Republic.

Dagan, I., Glickman, O., Magnini, B. (2005). The PASCAL Recognizing Textual Entailment Challenge. In *Proc. of the First PASCAL Challenges Workshop on RTE.* Southampton, U.K.

De Marneffe, M.C., Rafferty, A. N., Manning C. D. (2008). Finding contradictions in text. *In Proc. of ACL-08.* Columbus, OH.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Garoufi, K. (2007). *Towards a Better Understanding of Applied Textual Entailment.* Master Thesis. Saarland University, Saarbrücken, Germany.

Kirk R., (2009). Building an annotated textual inference corpus for motion and space, In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009*, Singapore.

Magnini, B., Cabrio E. (2009). Combining Specialized Entailment Engines. In *Proceedings of LTC '09.* Poznan, Poland.

Szpektor, I., Shnarch, E., Dagan I. (2007). Instance-based Evaluation of Entailment Rule Acquisition. In *Proceedings of ACL-07.* Prague, Czech Republic.

Vanderwende, L., Coughlin, D., and Dolan, B. (2005). What Syntax can Contribute in Entailment Task. *In Proceedings of the First PASCAL Challenges Workshop on RTE.* Southampton, U.K.