Contradiction-Focused Qualitative Evaluation of Textual Entailment

Bernardo Magnini
FBK-Irst
Trento, Italy
magnini@fbk.eu

Elena Cabrio
FBK-Irst, University of Trento
Trento, Italy
cabrio@fbk.eu

Abstract

In this paper we investigate the relation between positive and negative pairs in Textual Entailment (TE), in order to highlight the role of contradiction in TE datasets. We base our analysis on the decomposition of Text-Hypothesis pairs into *monothematic pairs*, i.e. pairs where only one linguistic phenomenon at a time is responsible for entailment judgment and we argue that such a deeper inspection of the linguistic phenomena behind textual entailment is necessary in order to highlight the role of contradiction. We support our analysis with a number of empirical experiments, which use current available TE systems.

1 Introduction

Textual Entailment (TE) (Dagan et al., 2009) provides a powerful and general framework for applied semantics. TE has been exploited in a series of evaluation campaigns (RTE - Recognizing Textual Entailment) (Bentivogli et al., 2009), where systems are asked to automatically judge whether the meaning of a portion of text, referred as Text (*T*), entails the meaning of another text, referred as Hypothesis (*H*).

RTE datasets have been mainly built with the purpose of showing the applicability of the TE framework to different semantic applications in Computational Linguistics. Starting from 2005, *[T,H]* pairs were created including samples from summarization, question answering, information extraction, and other applications. This evaluation provides useful cues for researchers and developers aiming at the integration of TE components in larger applications (see, for instance, the use of a TE engine for question answering in the QALL-ME project system[1], the use in relation extraction (Romano et al., 2006), and in reading comprehension systems (Nielsen et al., 2009)).

Although the RTE evaluations showed progresses in TE technologies, we think that there is still large room for improving qualitative analysis of both the RTE datasets and the system results. In particular, we intend to focus this paper on contradiction judgments and on a deep inspection of the linguistic phenomena that determine such judgments. More specifically, we address two distinguishing aspects of TE: (i) the variety of linguistic phenomena that are relevant for contradiction and how their distribution is represented in RTE datasets; (ii) the fact that in TE it is not enough to detect the polarity of a sentence, as in traditional semantic analysis, but rather it is necessary to analyze the dependencies between two sentences (i.e. the *[T,H]* pair) in order to establish whether a contradiction holds between the pair. Under this respect we are interested to investigate both how polarity among Text and Hypothesis affects the entailment/contradiction judgments and how different linguistic phenomena interact with polarity (e.g. whether specific combinations of phenomena are more frequent than others).

As an example, let us consider the pair:

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.

In order to detect the correct contradiction judgment between T and H it is necessary to solve the semantic inference that being the new President of a country is not compatible with being the outgoing President of the same country. This kind of inference requires that (i) the semantic opposition is detected, and that (ii) such opposition is consid-

[1] http://qallme.fbk.eu/

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 86–94,
Uppsala, July 2010.

		Text snippet (pair 125)	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. [...]	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. [...]	syntactic:argument-realization	E
	H3	**Felipe Calderon** is Mexico's **new president**.	syntactic:apposition	E

Table 1: Application of the decomposition methodology to an original RTE pair

ered relevant for the contradiction judgment in the specific context of the pair.

In order to address the issues above, we propose a methodology based on the decomposition of *[T,H]* pairs into *monothematic pairs*, each representing one single linguistic phenomenon relevant for entailment judgment. Then, the analysis is carried out both on the original *[T,H]* pair and on the monothematic pairs originated from it. In particular, we investigate the correlations on positive and on negative pairs separately, and we show that the strategies adopted by the TE systems to deal with phenomena contributing to the entailment or to the contradiction judgment come to light when analyzed using qualitative criteria. We have experimented the decomposition methodology over a dataset of pairs, which either are marked with a contradiction judgment, or show a polarity phenomenon (either in T or H) which, although present, is not relevant for cotradiction.

The final goal underlying our analysis of contradiction in current RTE datasets is to discover good strategies for systems to manage contradiction and, more generally, entailment judgments. To this aim, in Section 5 we propose a comparison between two systems participating at the last RTE-5 campaign and try to analyze their behaviour according to the decomposition into monothematic pairs.

The paper is structured as follows. Section 2 presents the main aspects related to contradiction within the RTE context. Section 3 explains the procedure for the creation of monothematic pairs starting from RTE pairs. Section 4 describes the experimental setup of our pilot study, as well as the results of the qualitative analysis. Section 5 outlines the preliminary achievements in terms of comparison of systems' strategies in order to manage contradiction. Finally, Section 6 reports on previous work on contradiction and textual entailment.

2 Contradiction and Textual Entailment

In RTE, two kinds of judgment are allowed: two ways (*yes* or *no* entailment) or three way judgment. In the latter, systems are required to decide whether the hypothesis is entailed by the text (*entailment*), contradicts the text (*contradiction*), or is neither entailed by nor contradicts the text (*unknown*). The RTE-4 and RTE-5 datasets are annotated for a 3-way decision: entailment (50% of the pairs), unknown (35%), contradiction (15%). This distribution among the three entailment judgments aims at reflecting the natural distribution of entailment in a corpus, where the percentage of text snippets neither entailing nor contradicting each other is higher than the contradicting ones. Even if this balance seems artificial since in a natural setting the presence of unknown pairs is much higher than the other two judgments (as demonstrated in the Pilot Task proposed in RTE-5 (Bentivogli et al., 2009)), the reason behind the choice of RTE organizers is to maintain a trade-off between the natural distribution of the data in real documents, and the creation of a dataset balanced beween positive and negative examples (as in two way task).

As already pointed out in (Wang, 2009), the similarity between T's and H's in pairs marked as entailment and contradiction is much higher with respect to the similarity between T's and H's in pairs marked as unknown. To support this intuition, (Bentivogli et al., 2009) provides some data on the lexical overlap between T's and H's in the last RTE Challenges. For instance, in RTE-4 the lexical overlap is 68.95% in entailment pairs, 67.97% in contradiction pairs and only 57.36% in the unknown pairs. Similarly, in RTE-5 the lexical overlap between T's and H's is 77.14% in entailment pairs, 78.93% in contradiction pairs and only 62.28% in the unknown pairs.

For this reason, for contradiction detection it is not sufficient to highlight mismatching information between sentences, but deeper comprehension is required. For applications in information analysis, it can be very important to detect incompatibility and discrepancies in the description of the same event, and the contradiction judgment in the TE task aims at covering this aspect. More specifically, in the RTE task the contradiction judgment is assigned to a T,H pair when the two text fragments are extremely unlikely to be true simultaneously.

According to Marneffe *et al.* (2008), contradictions may arise from a number of different constructions, defined in two primary categories: *i)* those occurring via antonymy, negation, and numeric mismatch, and *ii)* contradictions arising from the use of factive or modal words, structural and subtle lexical contrasts, and world knowledge. Comparing the distribution of contradiction types for RTE-3 and the real contradiction corpus they created collecting contradiction "in the wild" (e.g. from newswire, Wikipedia), they noticed that in the latter there is a much higher rate of negations, numeric and lexical contradictions with respect to RTE dataset, where contradictions of category *(ii)* occur more frequently. Analyzing RTE data of the previous challenges, we noticed that the tendency towards longer and more complex sentences in the datasets in order to reproduce more realistic scenarios, is also reflected in more complex structures determining contradictions. For instance, contradictions arising from overt negation as in (pair 1663, RTE-1 test set):

T: All residential areas in South Africa are segregated by race and no black neighborhoods have been established in Port Nolloth.

H: Black neighborhoods are located in Port Nolloth.

are infrequent in the datasets of more recent RTE challenges. For instance, in RTE-5 test set, only in 4 out of 90 contradiction pairs an overt negation is responsible for the contradiction judgment. In agreement with (Marneffe et al., 2008), we also remarked that most of the contradiction involve numeric mismatch, wrong appositions, entity mismatch and, above all, deeper inferences depending on background and world knowledge, as in (pair 567, RTE-5 test set):

T: "[...] we've done a series of tests on Senator Kennedy to determine the cause of his seizure. He has had no further seizures, remains in good overall condition, and is up and walking around the hospital".

H: Ted Kennedy is dead.

These considerations do not mean that overt negations do not appear in the RTE pairs. On the contrary, they are often present in T,H pairs, but most of the times their presence is irrelevant in the assignment of the correct entailment judgment to the pair. For instance, the scope of the negation can be a phrase or a sentence with additional information with respect to the relevant parts of T and H that allow to correctly judge the pair. This fact could be misleading for systems that do not correcly exploit syntactic information, as the experiments using Linear Distance described in (Cabrio et al., 2008).

3 Decomposing RTE pairs

The qualitative evaluation we propose takes advantage of previous work on monothematic datasets. A *monothematic pair* (Magnini and Cabrio, 2009) is defined as a *[T,H]* pair in which a certain phenomenon relevant to the entailment relation is highlighted and isolated. The main idea is to create such monothematic pairs on the basis of the phenomena which are actually present in the original RTE pairs, so that the actual distribution of the linguistic phenomena involved in the entailment relation emerges.

For the decomposition procedure, we refer to the methodology described in (Bentivogli et al., 2010), consisting of a number of steps carried out manually. The starting point is a *[T,H]* pair taken from one of the RTE datasets, that should be decomposed in a number of monothematic pairs $[T, H_i]_{mono}$, where T is the original Text and H_i are the Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in *[T,H]*.

In detail, the procedure for the creation of monothematic pairs is composed of 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 left hand side (LHS) of the rule, and information from H on *i* as the right hand side (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 $[T, H_i]_{mono}$. Mark the pair with phenomenon *i*.

3. Assign an entailment judgment to each monothematic pair.

Relevant linguistic phenomena are grouped using both fine-grained categories and broader categories. 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 (Table 2 reports a list of some of the phenomena detected in RTE-5 dataset).

Table 1 shows an example of the decomposition of a RTE pair (marked as *contradiction*) into monothematic pairs. At step 1 of the methodology both the phenomena that preserve the entailment and the phenomena that break the entailment rules causing a contradiction in the pair are detected, i.e. argument realization, apposition and semantic opposition (column *phenomena* in the table). While the monothematic pairs created basing on the first two phenomena preserve the entailment, the semantic opposition generates a contradiction (column *judgment*).

As an example, let's apply step by step the procedure to the phenomenon of semantic opposition. At step 2a of the methodology the general rule:

Pattern: x $\Leftarrow$ / $\Rightarrow$ y
Constraint: *semantic opposition(y,x)*

is instantiated (*new*$\Leftarrow$ / $\Rightarrow$*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 (column *text snippet* in the table) and marked as *semantic opposition* (macro-category *lexical*), and the pair is judged as *contradiction*.

In (Bentivogli et al., 2010), critical issues concerning the application of such procedure are discussed in detail, and more examples are provided. Furthermore, a pilot resource is created, composed of a first dataset with 60 pairs from RTE-5 test set (30 *positive*, and 30 *negative* randomly extracted examples), and a dataset composed of all the monothematic pairs derived by the first one following the procedure described before. The second dataset is composed of 167 pairs (134 *entailment*, 33 *contradiction* examples, considering 35 different linguistic phenomena).[2]

4 Analysis and discussion

Our analysis has been carried out taking advantage of the pilot resource created by Bentivogli *et al.* (2010). From their first dataset we extracted a sample of 48 pairs ($[T, H]_{sample-contr}$) composed of 30 *contradiction* pairs and 18 *entailment* pairs, the latter containing either in T or in H a directly or an indirectly licensed negation.[3] Furthermore, a dataset of 129 monothematic pairs (96 *entailment* and 33 *contradiction* examples), i.e. $[T, H]_{mono-contr}$, was derived by the pairs in $[T, H]_{sample-contr}$ applying the procedure described in Section 3. The linguistic phenomena isolated in the monothematic pairs (i.e. considered relevant to correctly assign the entailment judgment to our sample) are listed in Table 2.

In RTE datasets only a subpart of the potentially problematic phenomena concerning negation and negative polarity items is represented. At the same time, the specificity of the task lies in the fact that it is not enough to find the correct representation of the linguistic phenomena underlying a sentence meaning, but correct inferences should be derived from the relations that these phenomena contribute to establish between two text fragments. The mere presence of a negation in T is not relevant for the TE task, unless the scope of the negation (a token or a phrase) is present as non-negated in H

[2] Both datasets are freely available at http://hlt.fbk.eu/en/Technology/TE_Specialized_Data

[3] Following (Harabagiu et al., 2006) overt (directly licensed) negations include *i)* overt negative markers such as *not, n't*; *ii)* negative quantifiers as *no*, and expressions such as *no one* and *nothing*; *iii)* strong negative adverbs like *never*. Indirectly licensed negations include: *i)* verbs or phrasal verbs (e.g. *deny*, *fail*, *refuse*, *keep from*); *ii)* prepositions (e.g. *without*, *except*); weak quantifiers (e.g. *few*, *any*, *some*), and *iv)* traditional negative polarity items (e.g. *a red cent* or *anymore*).

phenomena	**# pairs** $[T,H]_{RTE5-mono-contr}$			
	entailment		**contradiction**	
	# mono	**probab.**	**# mono**	**probab.**
lex:identity	1	0.25	3	0.75
lex:format	2	1	-	-
lex:acronymy	1	1	-	-
lex:demonymy	1	1	-	-
lex:synonymy	6	1	-	-
lex:semantic-opp.	-	-	3	1
lex:hypernymy	2	1	-	-
TOT lexical	13	0.68	6	0.32
lexsynt:transp-head	2	1	-	-
lexsynt:verb-nom.	6	1	-	-
lexsynt:causative	1	1	-	-
lexsynt:paraphrase	2	1	-	-
TOT lexical-syntactic	11	1	-	-
synt:negation	-	-	1	1
synt:modifier	3	0.75	1	0.25
synt:arg-realization	4	1	-	-
synt:apposition	9	0.6	6	0.4
synt:list	1	1	-	-
synt:coordination	2	1	-	-
synt:actpass-altern.	4	0.67	2	0.33
TOT syntactic	23	0.7	10	0.3
disc:coreference	16	1	-	-
disc:apposition	2	1	-	-
disc:anaphora-zero	3	1	-	-
disc:ellipsis	3	1	-	-
disc:statements	1	1	-	-
TOT discourse	25	1	-	-
reas:apposition	1	0.5	1	0.5
reas:modifier	2	1	-	-
reas:genitive	1	1	-	-
reas:meronymy	1	0.5	1	0.5
reas:quantity	-	-	5	1
reas:spatial	1	1	-	-
reas:gen-inference	18	0.64	10	0.36
TOT reasoning	24	0.59	17	0.41
TOT (all phenomena)	96	0.74	33	0.26

Table 2: Occurrences of linguistic phenomena in TE contradiction pairs

(or viceversa), hence a contradiction is generated. For this reason, 18 pairs of $[T,H]_{sample-contr}$ are judged as *entailment* even if a negation is present, but it is not relevant to correctly assign the entailment judgment to the pair as in (pair 205, RTE-5 test set):

> *T: A team of European and American astronomers say that a recently discovered extrasolar planet, located* ***not*** *far from Earth, contains oceans and rivers of hot solid water. The team discovered the planet, Gliese 436 b [...].*
>
> *H: Gliese 436 b was found by scientists from America and Europe.*

As showed in Table 2, only in one pair of our sample the presence of a negation is relevant to assign the *contradiction* judgment to the pair. In the pairs we analyzed, contradiction mainly arise from quantity mismatching, semantic opposition (antonymy), mismatching appositions (e.g. *the Swiss Foreign Minister x* contradicts *y is the Swiss Foreign Minister*), and from general inference (e.g. *x became a naturalized citizen of the U.S.* contradicts *x is born in the U.S.*). Due to the small sample we analyzed, some phenomena appear rarely, and their distribution can not be considered as representative of the same phenomenon in a natural setting. In 27 out of 30 contradiction pairs, only one monothematic pair among the ones derived from each example was marked as contradiction, meaning that on average only one linguistic phenomenon is responsible for the contradiction judgment in a TE original pair. Hence the importance of detecting it.

Given the list of the phenomena isolated in $[T,H]_{mono-contr}$ with their frequency both in monothematic positive pairs and monothematic negative pairs, we derived the probability of linguistic phenomena to contribute more to the assignment of a certain judgment than to another (column *probab.* in Table 2). Such probability *P* of a phenomenon *i* to appear in a positive (or in a negative) pair is calculated as follows:

$$P(i|[T,H]_{positive}) = \frac{\#(i|[T,H]_{RTE5-positive-mono})}{\#(i|[T,H]_{RTE5-mono})} \qquad (1)$$

For instance, if the phenomenon *semantic opposition* appears in 3 pairs of our sample and all these pairs are marked as *contradiction*, we assign a probability of 1 to a pair containing a semantic opposition to be marked as contradiction. If the phenomenon *apposition* (syntax) appears in 9 monothematic positive pairs and in 6 negative pairs, that phenomenon has a probability of 0.6 to appear in positive examples and 0.4 to appear in negative examples. Due to their nature, some phenomena are strongly related to a certain judgment (e.g. semantic opposition), while other can appear both in positive and in negative pairs. Learning such correlations on larger datasets could be an interesting feature to be exploited by TE systems in the assignment of a certain judgment if the phenomenon i is detected in the pair.

Table 3 reports the cooccurrences of the linguistic phenomena relevant to inference in the pairs marked as *contradiction*. On the first horizontal row all the phenomena that at least in one pair determine contradiction are listed, while in the first column there are all the phenomena cooccurring with them in the pairs. The idea undelying this table is to understand if it is possible to identify recurrent patterns of cooccurrences between phenomena in contradiction pairs. As can be noticed, almost all phenomena occur together with expressions requiring deeper inference

	lex:identity	lex:sem_opposition	synt:negation	synt:modifier	synt:apposition	synt:actpass_altern	reas:meronymy	reas:quantity	reas:gen_inference
lex:identity								1	1
lex:format								1	
lex:acronymy					1				
lex:synonymy	1					1		1	1
lex:hypernymy								1	
lexsynt:vrb-nom	1	1						1	
lexsynt:caus.								1	
synt:modifier									1
synt:arg-realiz.		1							1
synt:apposition		2							3
synt:coord.								1	
synt:actpass	1	1							
disc:coref.	3				1				4
disc:apposition									
disc:anaph-0							1	1	
disc:ellipsis	1	1							2
disc:statements									1
reas:genitive			1						
reas:meronymy								1	
reas:gen-infer.	1			1	3		1	2	1

Table 3: Cooccurrencies of phenomena in contradiction pairs

(*reas:general_inference*), but this is due to the fact that this category is the most frequent one. Beside this, it seems that no specific patterns can be highlighted, but it could be worth to extend this analysis increasing the number of pairs of the sample.

5 Comparing RTE systems' behaviour on contradiction pairs

As introduced before, from a contradiction pair it is possible to extract on average 3 monothematic pairs (Bentivogli et al., 2009), and only one of these monothematic pairs is marked as contradiction. This means that on average only one linguistic phenomenon is responsible for the contradiction judgment in a RTE pair, while the others maintain the entailment relation (i.e. it is possible to correcly apply an entailment rule as exemplified in Section 3). On the contrary, in a pair judged as entailment, all the monothematic pairs derived from it are marked as entailment.

These observations point out the fact that if a TE system is able to correctly isolate and judge the phenomenon that generates the contradiction, the system should be able to assign the correct judgment to the original contradiction pair, despite possible mistakes in handling the other phenomena present in that pair.

In order to understand how it is possible to take advantage of the data analyzed so far to improve a TE system, we run two systems that took part into the last RTE challenge (RTE-5) on $[T,H]_{mono-contr}$.

The first system we used is the EDITS system (Edit Distance Textual Entailment Suite) (Negri et al., 2009)[4], that assumes that the distance between *T* and *H* is a characteristics that separates the positive pairs, for which entailment holds, from the negative pairs, for which entailment does not hold (it is developed according to the two way task). It is based on edit distance algorithms, and computes the *[T,H]* distance as the overall cost of the edit operations (i.e. *insertion*, *deletion* and *substitution*) that are required to transform *T* into *H*. In particular, we applied the model that produced EDITS best run at RTE-5 (acc. on RTE-5 test set: 60.2%). The main features of this run are: Tree Edit Distance algorithm on the parsed trees of *T* and *H*, Wikipedia lexical entailment rules, and PSO optimized operation costs, as described in (Mehdad et al., 2009).

The other system used in our experiments is VENSES[5] (Delmonte et al., 2009), that obtained performances similar to EDITS at RTE-5 (acc. on test set: 61.5%). VENSES applies a linguistically-based approach for semantic inference, composed of two main components: *i)* a grammatically-driven subsystem that validates the well-formedness of the predicate-argument structure and works on the output of a deep parser producing augmented (i.e. fully indexed) head-dependency structures; and *ii)* a subsystem that detects allowed logical and lexical inferences basing on different kind of structural transformations intended to produce a semantically valid meaning correspondence. The system has a pronominal binding module that works at text/hypothesis level separately for lexical personal, possessive and reflexive pronouns, which are substituted by the heads of their antecedents. Also in this case, we applied the same configuration of the system used in RTE evaluation.

Table 4 reports EDITS and VENSES accuracies on the monothematic pairs of $[T,H]_{mono-contr}$.

As said before, the accuracy reported for some very rare phenomena cannot be considered completely reliable. Nevertheless, from these data the main features of the systems can be identified. For instance, EDITS obtains the highest accuracies on the positive monothematic pairs, while it seems it has no peculiar strategies to deal with phenomena

[4] http://edits.fbk.eu/

[5] http://project.cgm.unive.it/venses_en.html

phenomena	EDITS % acc.		VENSES % acc.	
	pos.	neg.	pos.	neg.
lex:identity	100	0	100	33.3
lex:format	100	-	100	-
lex:acronymy	100	-	0	-
lex:demonymy	100	-	100	-
lex:synonymy	80.3	-	80.3	-
lex:semantic-opp.	-	0	-	100
lex:hypernymy	100	-	100	-
TOT lexical	96.7	0	80	66.6
lexsynt:transp-head	100	-	50	-
lexsynt:verb-nom.	83.3	-	16	-
lexsynt:causative	100	-	100	-
lexsynt:paraphrase	100	-	100	-
TOT lexical-syntactic	95.8	-	66.5	-
synt:negation	-	0	-	0
synt:modifier	100	0	33.3	100
synt:arg-realization	100	-	50	-
synt:apposition	100	33.3	55.5	83.3
synt:list	100	-	100	-
synt:coordination	100	-	50	-
synt:actpass-altern.	100	0	25	50
TOT syntactic	100	22.2	52.3	77.7
disc:coreference	95	-	50	-
disc:apposition	100	-	0	-
disc:anaphora-zero	100	-	33.3	-
disc:ellipsis	100	-	33.3	-
disc:statements	100	-	0	-
TOT discourse	99	-	23.3	-
reas:apposition	100	0	100	100
reas:modifier	50	-	100	-
reas:genitive	100	-	100	-
reas:meronymy	100	0	100	0
reas:quantity	-	0	-	80
reas:spatial	100	-	0	-
reas:gen-inference	87.5	50	37.5	90
TOT reasoning	89.5	35.2	72.9	82.3
TOT (all phenomena)	96.2	25	59	81.2

Table 4: RTE systems' accuracy on phenomena

that generally cause contradiction (e.g. *semantic opposition*, *negation*, and *quantity mismatching*). On the contrary, VENSES shows an opposite behaviour, obtaining the best results on the negative cases. Analysing such data it is possible to hypothesize systems' behaviours: for example, on the monothematic dataset EDITS produces a pretty high number of false positives, meaning that for this system if there are no evidences of contradiction, a pair should be marked as entailment (in order to improve such system, strategies to detect contradiction pairs should be thought). On the contrary, VENSES produces a pretty high number of false negatives, meaning that if the system is not able to find evidences of entailment, it assigns the contradiction value to the pairs (for this system, being able to correctly detect all the phenomena contributing to entailment in a pair is fundamental, otherwise it will be marked as contradiction).

6 Related Work

Condoravdi *et al.* (2003) first proposed contradiction detection as an important NLP task, then (Harabagiu et al., 2006) provided the first empirical results for it, focusing on contradiction caused by negation, antonymy, and paraphrases. Voorhees (2008) carries out an analysis of RTE-3 extended task, examining systems' abilities to detect contradiction and providing explanations of their reasoning when making entailment decisions.

Beside defining the categories of construction from which contradiction may arise, Marneffe *et al.* (2008) provide the annotation of the RTE datasets (RTE-1 and RTE-2) for contradiction. Furthermore, they also collect contradiction "in the wild" (e.g. from newswire, Wikipedia) to sample naturally occurring ones.[6]

Ritter *et al.* (2008) extend (Marneffe et al., 2008)'s analysis to a class of contradiction that can only be detected using backgroud knowledge, and describe a case study of contradiction detection based on functional relations. They also automatically generate a corpus of seeming contradiction from the Web text.[7]

Furthermore, some of the systems presented in the previous editions of the RTE challenges attempted specic strategies to focus on the phenomenon of negation. For instance, (Snow et al., 2006) presents a framework for recognizing textual entailment that focuses on the use of syntactic heuristics to recognize false entailment. Among the others, heuristics concerning negation mismatch and antonym match are defined. In (Tatu et al., 2007) the logic representation of sentences with negated concepts was altered to mark as negated the entire scope of the negation. (Ferrandez et al., 2009) propose a system facing the entailment recognition by computing shallow lexical deductions and richer inferences based on semantics, and features relating to negation are extracted. In (Iftene et al., 2009) several rules are extracted and applied to detect contradiction cases.

7 Conclusion

We have proposed a methodology for the qualitative analysis of TE systems focusing on contradiction judgments and on the linguistic phenomena that determine such judgments. The methodology is based on the decomposition of *[T,H]* pairs into *monothematic pairs*, each representing one single linguistic phenomenon relevant for entailment

[6] Their corpora are available at http://www-nlp.stanford.edu/projects/contradiction.

[7] Available at http://www.cs.washington.edu/research/ aucontraire/

judgment.

In particular, the phenomena from which contradiction may arise and their distribution in RTE datasets have been highlighted, and a pilot study comparing the performancies of two RTE systems both on monothematic pairs and on the corresponding original ones has been carried out. We discovered that, although the two systems have similar performances in terms of accuracy on the RTE-5 datasets, they show significant differences in their respective abilities to correctly manage different linguistic phenomena that generally cause contradiction. We hope that the analysis of contradiction in current RTE datasets may bring interesting elements to TE system developers to define good strategies to manage contradiction and, more generally, entailment judgments.

8 Acknowledgements

This work has been partially supported by the LiveMemories project (Active Digital Memories of Collective Life) funded by the Autonomous Province of Trento under the call "Major Projects". We would like to thank Professor Rodolfo Delmonte and Sara Tonelli for running the VENSES system on our datasets.

References

Bentivogli, Luisa, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The Fifth PASCAL RTE Challenge. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. To appear. Gaithersburg, Maryland. 17 November.

Bentivogli, Luisa, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. *Proceedings of the 7th LREC conference*. Valletta, Malta. 19-21 May.

Cabrio, Elena, Milen Ognianov Kouylekov and Bernardo Magnini, 2008. Combining Specialized Entailment Engines for RTE-4, *Proceedings of the Text Analysis Conference (TAC 2008)*. Gaithersburg, Maryland, USA, 17-18 November.

Condoravdi, Cleo, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel Bobrow. 2003. Entailment, Intentionality and Text Understanding *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*. Edmonton, Alberta, Canada. 31 May.

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, Volume 15, Special Issue 04, October 2009, pp i-xvii. Cambridge University Press.

De Marneffe, Marie-Catherine, Anna N. Rafferty and Christopher D. Manning. 2008. Finding Contradictions in Text. *Proceedings of ACL-08: HLT*, pages 10391047. Columbus, Ohio, USA, June.

Delmonte, Rodolfo, Sara Tonelli, Rocco Tripodi. 2009. Semantic Processing for Text Entailment with VENSES. *Proceedings of the TAC 2009 Workshop on TE*. Gaithersburg, Maryland. 17 November.

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

Ferrández, Óscar, Rafael Muñoz, and Manuel Palomar. 2009. Alicante University at TAC 2009: Experiments in RTE. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. Gaithersburg, Maryland. 17 November.

Harabagiu, Sanda, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In Proceedings of AAAI-06. Boston, Massachusetts. July 16-20.

Iftene, Adrian, Mihai-Alex Moruz 2009. UAIC Participation at RTE-5. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. To appear. Gaithersburg, Maryland. 17 November.

Magnini, Bernardo, and Elena Cabrio. 2009. Combining Specialized Entailment Engines. *Proceedings of the LTC '09 conference*. Poznan, Poland. 6-8 November.

Mehdad, Yashar, Matteo Negri, Elena Cabrio, Milen Kouylekov, and Bernardo Magnini. 2009. Using Lexical Resources in a Distance-Based Approach to RTE. *Proceedings of the TAC 2009 Workshop on TE*. Gaithersburg, Maryland. 17 November 2009.

Negri, Matteo, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. 2009. Towards Extensible Textual Entailment Engines: the EDITS Package. *AI*IA 2009: Emergent Perspectives in Artificial Intelligence*. Lecture Notes in Computer Science, Springer-Verlag, pp. 314-323. 2009.

Nielsen, Rodney D., Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. In Ido Dagan, Bill Dolan, Bernardo Magnini and Dan Roth (Eds.) *The Journal of Natural Language Engineering, (JNLE)*. 15, pp 479-501. Copyright Cambridge University Press, Cambridge, United Kingdom.

Ritter, Alan, Doug Downey, Stephen Soderland, and Oren Etzioni. 2008. It's a Contradiction - No, it's not: A Case Study using Functional Relations. *Proceedings of 2008 Conference on Empirical Methods* *in Natural Language Processing*. Honolulu, Hawaii. 25-27 October.

Romano, Lorenza, Milen Ognianov Kouylekov, Idan Szpektor, Ido Kalman Dagan, and Alberto Lavelli. 2006. Investigating a Generic Paraphrase-Based Approach for Relation Extraction. *Proceedings of EACL 2006*. Trento, Italy. 3-7 April.

Snow, Rion, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.* New York, 4-9 June.

Tatu, Marta, Dan I. Moldovan. 2007. COGEX at RTE 3. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.* Prague, Czech Republic, 28-29 June.

Voorhees, Ellen M. 2008. Contradictions and Justifications: Extentions to the Textual Entailment Task. *Proceedings of ACL-08: HLT*. Columbus, Ohio, USA. 15-20 June.

Wang, Rui, and Yi Zhang. 2009. Recognizing Textual Relatedness with Predicate-Argument Structures. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 6-7 August.