Toward Qualitative Evaluation of Textual Entailment Systems

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

Bernardo Magnini
FBK-Irst
magnini@fbk.eu

Abstract

This paper presents a methodology for a quantitative and qualitative evaluation of Textual Entailment systems. We take advantage of 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 propose to run TE systems over such datasets. We show that several behaviours of a system can be explained in terms of the correlation between the accuracy on monothematic pairs and the accuracy on the corresponding original pairs.

1 Introduction

Since 2005, Recognizing Textual Entailment (RTE) has been proposed as a task whose aim is to capture major semantic inference needs across applications in Computational Linguistics (Dagan et al., 2009). 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*). 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 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 the following aspects:

1. There is relatively poor analysis of the linguistic phenomena that are relevant for the RTE datasets, and very little is known about the distribution of such phenomena, and about the ability of participating systems to correctly detect and judge them in *T,H* pairs. Experiments like the ablation tests attempted in the last RTE-5 campaign on lexical and lexical-syntactic resources go in this direction, although the degree of comprehension is still far from being optimal.

2. We are interested in the correlations among the capability of a system to address single linguistic phenomena in a pair and the ability to correctly judge the pair itself. Despite the strong intuition about such correlation (i.e. the more the phenomena for which a system is trained, the better the final judgment), no empirical evidences support it.

3. Although the ability to detect and manage single phenomena seems to be a crucial feature of high performing systems, very little is known about how systems manage to combine such results in a global score for a pair. The mechanism underlying such composition may shed light on meaning composition related to TE tasks.

4. Finally, we are interested in the relation between the above mentioned items over the different kinds of pairs represented in RTE

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

Coling 2008: Poster Volume, pages 99–107,
Beijing, August 2010

datasets, specifically *entailment*, *contradiction* and *unknown* pairs. In this case the intuition is that some phenomena are more relevant for a certain judgment rather than for another.

To address the issues above, we propose an evaluation methodology aiming at providing a number of quantitative and qualitative indicators about a TE system. The method is based on the decomposition of *T,H* pairs into *monothematic pairs*, each representing one single linguistic phenomenon relevant for entailment judgment. Evaluation is carried out both on the original *T,H* pair and on the monothematic pairs originated from it. We define a correlation index between the accuracy of the system on the original *T,H* pairs and the accuracy on the corresponding monothematic pairs. We investigate the use of such correlations on different subsets of the evaluation dataset (i.e. positive vs negative pairs) and we try to induce regular patterns of evaluation.

The method we propose has been tested on a sample of 60 pairs, each decomposed in the corresponding monothematic pairs, and using two systems that obtained similar performances in RTE-5. We show that the main features and differences of these systems come to light when evaluated using qualitative criteria. Futhermore, we compare such systems with two different baseline systems, the first one performing Word Overlap, while the second one is an ideal system that knows *a priori* the probability of a linguistic phenomenon to be associated with a certain entailment judgement.

The paper is structured as follows. Section 2 explains the procedure for the creation of monothematic pairs starting from RTE pairs. Section 3 presents the evaluation methodology we propose, while Section 4 describes our pilot study. Section 5 concludes the paper and proposes future developments.

2 Decomposing RTE pairs

Our proposal on qualitative evaluation takes advantage of previous work on specialized entailment engines and monothematic datasets. A *monothematic pair* is defined (Magnini and Cabrio, 2009) 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 the monothematic pairs basing on the phenomena that 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 data sets, that should be decomposed in a number of monothematic pairs $[T, H_i]$, 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 details, the procedure for the creation of monothematic pairs is composed of the following steps:

1. Individuate the phenomena contributing to the entailment decision in *[T,H]*.

2. For each linguistic phenomenon *i*:

   (a) Detect a general entailment rule r_i for *i*, and instantiate it using the part of T expressing *i* as the left hand side (LHS) of the rule, and information from H on *i* as the right side (RHS).

   (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.

Relevant linguistic phenomena are grouped using both fine-grained categories and broader categories, defined referring to widely accepted classifications 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 lists the phenomena detected in RTE-5 datasets).

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 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

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 those 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*.

3 Evaluation methodology

Aim of the evaluation methodology we propose is to provide quantitative and qualitative indicators about the behaviours of actual TE systems.

3.1 General Method

The basic assumption of the evaluation methodology is that the more a system is able to correctly solve the linguistic phenomena underlying the entailment relation separately, the more the system should be able to correctly judge more complex pairs, in which different phenomena are present and interact in a complex way. Such assumption is motivated by the notion of meaning compositionality, according to which the meaning of a complex expression *e* in a language *L* is determined by the structure of *e* in *L* and the meaning of the constituents of *e* in *L* (Frege, 1892). In a parallel way, we assume that it is possible to understand the entailment relation of a *T,H* pair (i.e. to correctly judge the *entailment/contradiction* relation) only if all the phenomena contributing to such relation are solved.

According to such assumption, we expect that the higher the accuracy of a system on the monothematic pairs and the compositional strategy, the better its performances on the original RTE pairs. Furthermore, the precision a system gains on single phenomena should be maintained over the general dataset, thanks to suitable mechanisms of meaning combination.

Given a dataset composed of original RTE pairs $[T, H]$, a dataset composed of all the monothematic pairs derived from it $[T, H]_{mono}$, and a TE system *S*, the evaluation methodology we propose consists of the following steps:

1. Run *S* both on $[T, H]$ and on $[T, H]_{mono}$, to obtain the accuracies of *S* both on the RTE original and on monothematic pairs;

2. Extract data concerning the behaviour of *S* on each phenomenon or on categories of phenomena, and calculate separate accuracies. This way it is possible to evaluate how much a system is able to correctly deal with single or with categories of phenomena;

3. Calculate the correlation between the ability of the system to correctly judge the monothematic pairs of $[T, H]_{mono}$ with respect to the

ability to correctly judge the original ones in $[T, H]$. Such correlation is expressed through a *Correlation Index (CI)*, as defined in Section 3.2;

4. In order to check if the same *CI* is maintained over both entailment and contradiction pairs (i.e. to verify if the system has peculiar strategies to correctly assign both judgments, and if the high similarity of monothematic pairs does not bias its behaviour), we calculate a *Deviation Index (DI)* as the difference between the *CIs* on entailment and on contradiction pairs, as explained in more details in Section 3.3.

3.2 Correlation Index (CI)

As introduced before, we assume that the accuracy obtained on $[T, H]_{mono}$ should positively correlate with the accuracy obtained on $[T, H]$. We define a *Correlation Index* as the ratio between the accuracy of the system on the original RTE dataset and the accuracy obtained on the monothematic dataset, as follows:

$$CI = \frac{acc[T, H]}{acc[T, H]_{mono}} \quad (1)$$

We expect the correlation index of an optimal ideal system (or the human goldstandard) to be equal to 1, i.e. 100% accuracy on the monothematic dataset should correspond to 100% accuracy on the original RTE dataset. For this reason, we consider $CI = 1$ as the ideal correlation, and we calculate the difference between such ideal CI and the correlation obtained for a system *S*.

Given such expectations, CI_S can assume three different configurations with respect to the upperbound (i.e. the ideal correlation):

- $CI_S \cong 1$ (ideal correlation): When CI_S approaches to 1, the system shows high correlation with the ideal behaviour assumed by the compositionality principle. As a consequence, we can predict that improving single modules will correspondingly affect the global performance.

- $CI_S < 1$ (missing correlation): The system is not able to exploit the ability in solving single phenomena to correctly judge the original RTE pairs. This may be due to the fact that the system does not adopt suitable combination mechanisms and loses the potentiality shown by its performances on monothematic pairs.

- $CI_S > 1$ (over correlation): The system does not exploit the ability to solve single linguistic components to solve the whole pairs, and has different mechanisms to evaluate the entailment. Probably, such a system is not intended to be modularized.

Beside this "global" correlation index calculated on the complete RTE data and on all the monothematic pairs created from it, the *CI* can also be calculated *i)* on categories of phenomena, to verify which phenomena a system is more able to solve both when isolated and when interacting with other phenomena, e.g. :

$$CI_{lex} = \frac{acc[T, H]_{lex}}{acc[T, H]_{mono-lex}} \quad (2)$$

including in $[T, H]_{lex}$ all the pairs in which at least one lexical phenomenon is present and contribute to the entailment/contradiction judgments, and in $[T, H]_{mono-lex}$ all the monothematic pairs in which a lexical phenomenon is isolated; or *ii)* on kind of judgment (*entailment*, *contradiction*, *unknown*), allowing deeper qualitative analysis of the performances of a system.

3.3 Deviation Index (DI)

We explained that a low *CI* (i.e. < 1) of a system reflects the inability to correctly exploit the potentially promising results obtained on monothematic pairs to correctly judge RTE pairs. Actually, it could also be the case that the system does not perform a correct combination because even the results got on the monothematic pairs were due to chance (e.g. a word overlap system performs well on monothematic pairs because of the high similarity between *T* and *H*, and not because it has linguistic strategies).

We detect such cases by decomposing the evaluation datasets, separating positive (i.e. *entailment*) from negative (i.e. *contradiction*, *unknown*) examples both in $[T, H]$ and in $[T, H]_{mono}$, and

independently run *S* on the new datasets. Then, we have more fine grained evaluation patterns through which we can analyze the system behaviour.

In the ideal case, we expect to have good correlation between the accuracy obtained on the monothematic pairs and the accuracy obtained on the original ones ($0 < CI_{pos} \leq 1$ and $0 < CI_{neg} \leq 1$). On the contrary, we expect that systems either without a clear composition strategy or without strong components on specific linguistic phenomena (e.g. a word overlap system), would show a significant difference of correlation on the different datasets. More specifically, situations of *inverse correlation* on the entailment and contradiction pairs (e.g. over correlation on contradiction pairs and missing correlation on entailment pairs) may reveal that the system itself is affected by the nature of the dataset (i.e. its behaviour is biased by the high similarity of $[T, H]_{mono}$), and weaknesses in the ability of solving phenomena that more frequently contribute to the assignment of a contradiction (or an entailment) judgment come to light.

We formalize such intuition defining a *Deviation Index (DI)* as the difference between the correlation indexes, respectively, on entailment and contradiction/unknown pairs, as follows:

$$|DI| = CI_{pos} - CI_{neg} \quad (3)$$

For instance, an high Deviation Index due to a missing correlation on positive entailment pairs and an over correlation for negative pairs, is interpreted as an evidence that the system has low accuracy on $[T, H]_{mono}$ - *T* and *H* are very similar and the system has no strategies to understand that the phenomenon that is present has to be judged as contradictory -, and a higher accuracy on $[T, H]$, probably due to chance. In the ideal case $DI_S \cong 0$, since we assumed the ideal *CIs* on both positive and negative examples to be as close as possible to 1 (see Section 3.2).

4 Experiments and discussion

This Section describes the experimental setup of our pilot study, carried out using two systems that took part in RTE-5 i.e EDITS and VENSES. We show the results obtained and the qualitative analysis performed basing on the proposed evaluation methodology. Their respective *CIs* and *DIs* are compared with two baselines: a word overlap system, and a system biased by the knowledge of the probability that a linguistic phenomenon contributes to the assignment of a certain entailment judgment.

4.1 Dataset

The evaluation method has been tested on a dataset composed of 60 pairs from RTE-5 test set ($[T, H]_{RTE5-sample}$, composed of 30 *entailment*, and 30 *contradiction* randomly extracted examples), and a dataset composed of all the monothematic pairs derived by the first one following the procedure described in Section 2. The second dataset $[T, H]_{RTE5-mono}$ is composed of 167 pairs (135 *entailment*, 32 *contradiction* examples, considering 35 different linguistic phenomena)[2]. On average, 2.78 monothematic pairs have been created from the original pairs. In this pilot study we decided to limit our analysis to entailment and contradiction pairs since, as observed in (Bentivogli et al., 2010), in most of the unknown pairs no linguistic phenomena relating T to H could be detected.

4.2 TE systems

EDITS The EDITS system (Edit Distance Textual Entailment Suite)[3] (Negri et al., 2009) assumes that the distance between *T* and *H* is a characteristic that separates the positive pairs, for which entailment holds, from the negative pairs, for which entailment does not hold (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*) required to transform *T* into *H*. For our experiments we applied the model that produced EDITS best run at RTE-5 (acc. on test set: 60.2%). The main features are: Tree Edit Distance algorithm on the parsed trees of *T* and *H*, Wikipedia lexical entailment rules, and PSO optimized operation costs (Mehdad et al., 2009).

[2] http://hlt.fbk.eu/en/Technology/TE-_Specialized_Data

[3] Available as open source at http://edits.fbk.eu/

VENSES The other system used in our experiments is VENSES[4] (Delmonte et al., 2009), that obtained performances similar to EDITS at RTE-5 (acc. on test set: 61.5%). It applies a linguistically-based approach for semantic inference, and is composed of two main components: *i)* a grammatically-driven subsystem validates the well-formedness of the predicate-argument structure and works on the output of a deep parser producing augmented head-dependency structures; and *ii)* a subsystem detects allowed logical and lexical inferences basing on different kind of structural transformations intended to produce a semantically valid meaning correspondence. Also in this case, we applied the best configuration of the system used in RTE-5.

Baseline system 1: Word Overlap algorithm The first baseline applies a Word Overlap (WO) algorithm on tokenized text. The threshold to separate positive from negative pairs has been learnt on the whole RTE-5 training dataset.

Baseline system 2: Linguistic biased system The second baseline is produced by a more sophisticated but biased system. It exploits the probability of linguistic phenomena to contribute more to the assignment of a certain judgment than to another. Such probabilities are learnt on the $[T, H]_{RTE5-mono}$ goldstandard: given the list of the phenomena with their frequency in monothematic positive and negative pairs (columns 1,2,3 of Table 2), we calculate the probability *P* of phenomenon *i* to appear in a positive (or in a negative) pair as follows:

$$P(i|[T,H]_{positive}) = \frac{\#(i|[T,H]_{RTE5-positive-mono})}{\#(i|[T,H]_{RTE5-mono})} \quad (4)$$

For instance, if the phenomenon *apposition* appears in 11 monothematic positive pairs and in 6 negative pairs, it has a probability of 64.7% to appear in positive examples and 35.3% to appear in negative ones. Such knowledge is then stored in the system, and is used in the classification phase, assigning the most probable judgment associated to a certain phenomenon.

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

When applied to $[T, H]_{RTE5-sample}$, this system uses a simple combination strategy: if phenomena associated with different judgments are present in a pair, and one phenomenon is associated with a contradiction judgment with a probability $> 50\%$, the pair is marked as *contradiction*, otherwise it is marked as *entailment*.

4.3 Results

Following the methodology described in Section 3, at step 1 we run EDITS and VENSES on $[T, H]_{RTE5-sample}$, and on $[T, H]_{RTE5-mono}$ (Table 3 reports the accuracies obtained).

At step 2, we calculate the accuracy of EDITS and VENSES on each single linguistic phenomenon, and on categories of phenomena. Table 2 shows the distribution of the phenomena on the dataset, reflected in the number of positive and negative monothematic pairs created for each phenomenon. As can be seen, some phenomena appear more frequently than others (e.g. *coreference*, *general inference*). Furthermore, some linguistic phenomena allow only the creation of positive or negative examples, while others can contribute to the assignment of both judgments. Due to the small datasets we used, some phenomena appear rarely; the accuracy on them cannot be considered completely reliable.

Nevertheless, from these data the main features of the systems can be identified. For instance, EDITS obtains the highest accuracy on positive monothematic pairs, while it seems it has no peculiar strategies to deal with phenomena causing contradiction (e.g. *semantic opposition*, and *quantity mismatching*). On the contrary, VENSES shows an opposite behaviour, obtaining the best results on the negative cases.

At step 3 of the proposed evaluation methodology, we calculate the correlation index between the ability of the system to correctly judge the monothematic pairs of $[T, H]_{RTE5-mono}$ with respect to the ability to correctly judge the original ones in $[T, H]_{RTE5-sample}$.

Table 3 compares EDITS and VENSES *CI* with the two baseline systems described before. As can be noticed, even if EDITS *CI* outperforms the WO system, they show a similar behaviour (high accuracy on monothematic pairs, and much lower

phenomena	# $[T, H]$ $_{RTE5-mono}$		EDITS % acc.		VENSES % acc.	
	pos.	neg.	pos.	neg.	pos.	neg.
lex:identity	1	3	100	0	100	33.3
lex:format	2	-	100	-	100	-
lex:acronymy	3	-	100	-	33.3	-
lex:demonymy	1	-	100	-	100	-
lex:synonymy	11	-	90.9	-	90.9	-
lex:semantic-opp.	-	3	-	0	-	100
lex:hypernymy	3	-	100	-	66.6	-
lex:geo-knowledge	1	-	100	-	100	-
TOT lexical	22	6	95.4	0	77.2	66.6
lexsynt:transp-head	2	-	100	-	50	-
lexsynt:verb-nom.	8	-	87.5	-	25	-
lexsynt:causative	1	-	100	-	100	-
lexsynt:paraphrase	3	-	100	-	66.6	-
TOT lex-syntactic	14	-	92.8	-	42.8	-
synt:negation	-	1	-	0	-	0
synt:modifier	3	1	100	0	33.3	100
synt:arg-realization	5	-	100	-	40	-
synt:apposition	11	6	100	33.3	54.5	83.3
synt:list	1	-	100	-	100	-
synt:coordination	3	-	100	-	33.3	-
synt:actpass-altern.	4	2	100	0	25	50
TOT syntactic	28	9	96.4	22.2	42.8	77.7
disc:coreference	20	-	95	-	50	-
disc:apposition	3	-	100	-	0	-
disc:anaphora-zero	5	-	80	-	20	-
disc:ellipsis	4	-	100	-	25	-
disc:statements	1	-	100	-	0	-
TOT discourse	33	-	93.9	-	36.3	-
reas:apposition	2	1	100	0	50	100
reas:modifier	3	-	66.6	-	100	-
reas:genitive	1	-	100	-	100	-
reas:relative-clause	1	-	100	-	0	-
reas:elliptic-expr.	1	-	100	-	0	-
reas:meronymy	1	1	100	0	100	0
reas:metonymy	3	-	100	-	33.3	-
reas:representat.	1	-	100	-	0	-
reas:quantity	-	5	-	0	-	80
reas:spatial	1	-	100	-	0	-
reas:gen-inference	24	10	87.5	50	37.5	90
TOT reasoning	38	17	89.4	35.2	42.1	82.3
TOT (all phenom)	135	32	93.3	25	45.9	81.2

Table 2: Systems' accuracy on phenomena

on the RTE sample). According to our definition, their *CIs* ($0 < CI < 1$) show a good ability of the systems to deal with linguistic phenomena when isolated, but a scarce ability in combining them to assign the final judgment. EDITS *CI* is not far from the *CI* of the linguistic biased baseline system, even if we were expecting a higher *CI* for the latter system. The reason is that beside the linguistic phenomena that allow only the creation of negative monothematic pairs, all the phenomena that allow both judgments have a higher probability to contribute to the creation of positive monothematic pairs.

Comparing the *CI* of the four analyzed systems with the ideal correlation ($CI_S \cong 1$, see Section 3.2), VENSES is the closest one ($\Delta = 0.15$), even if it shows a light over correlation (probably due to the nature of the dataset). The second closest

	acc. % $_{RTE5-sample}$	acc. % $_{RTE5-mono}$	*CI*
EDITS	58.3	80.8	0.72
VENSES	60	52.6	1.15
Word Overlap	38.3	77.24	0.49
ling baseline	68.3	86.8	0.79

Table 3: Evaluation on RTE pairs and on monothematic pairs

		categories of linguistic phenomena				
	RTE5 data	lex.	lex-synt.	synt.	disc.	reas.
EDITS	$_{sample}$	47.8	64.3	51.7	75	62.5
	$_{mono}$	75	92.8	78.3	93.9	72.7
	CI	**0.63**	**0.69**	**0.66**	**0.79**	**0. 85**
VENSES	$_{sample}$	47.2	42.8	62	46.4	67.5
	$_{mono}$	75	42.8	51.3	33	54.5
	CI	**0.62**	**1**	**1.2**	**1.4**	**1.23**
WO baseline	$_{sample}$	36.3	57.1	34.4	50	35
	$_{mono}$	78.5	71.4	72.9	96.9	69
	CI	**0.46**	**0.79**	**0.47**	**0.51**	**0.5**
ling-biased baseline	$_{sample}$	82.6	92.8	58.6	82.1	70
	$_{mono}$	96.4	100	75.6	96.9	80
	CI	**0.85**	**0.92**	**0.77**	**0.84**	**0.87**

Table 4: Evaluation on categories of phenomena

one is the linguistic biased system ($\Delta = 0.21$), showing that the knowledge of the most probable judgment assigned to a certain phenomenon can be a useful information.

Table 4 reports an evaluation of the four systems on categories of linguistic phenomena.

To check if the same *CI* is maintained over both entailment and contradiction pairs, we calculate a *Deviation Index* as the difference between the *CIs* on entailment and on contradiction pairs (step 4 of our methodology). As described in Section 3, we created four datasets dividing both $[T, H]_{RTE5-sample}$ and $[T, H]_{RTE5-mono}$ into positive (i.e. *entailment*) and negative (i.e. *contradiction*) examples. We run EDITS and VENSES on the datasets and we calculate the *CI* on positive and on negative examples separately. If we obtained missing correlation between the accuracy on the monothematic pairs and the accuracy on RTE original ones, it would mean that the potentiality that the systems show on monothematic pairs is not exploited to correctly judge more complex pairs, therefore compositional mechanisms should be improved.

Table 5 shows that the *DIs* of the linguistic biased system and of VENSES are close to the ideal case ($DI_S \cong 0$), indicating a good capacity to correctly differentiate entailment from contradiction cases. EDITS results demonstrate that the

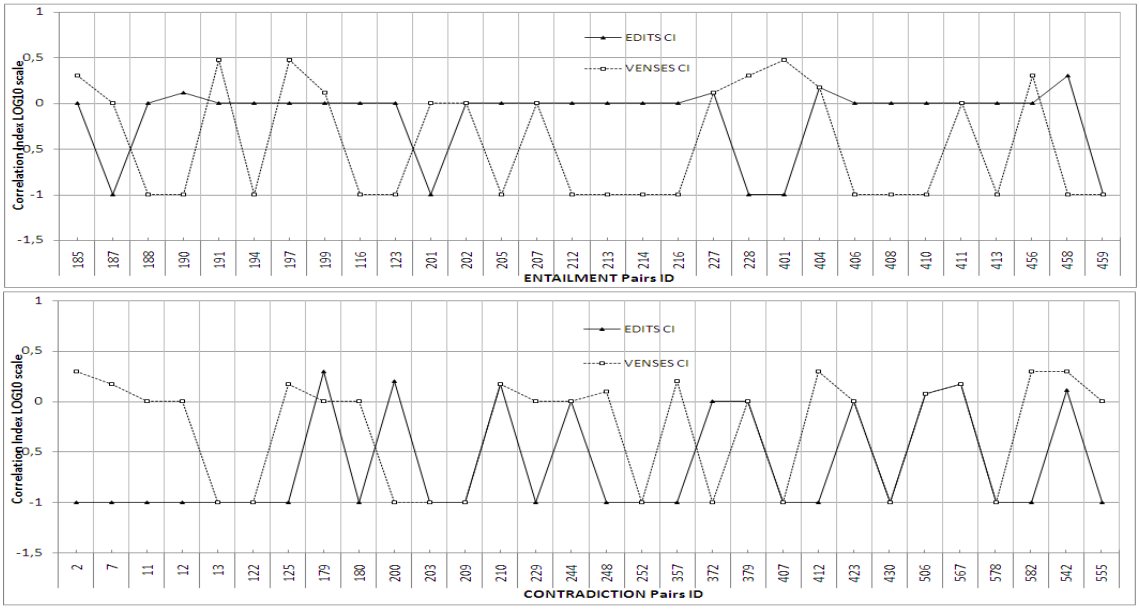

Figure 1: Correlation Index on entailment and contradiction pairs for EDITS and VENSES

		% acc. $_{RTE5\ sample}$	% acc. $_{RTE5\ mono}$	*CI*	*DI*
EDITS	E	83.3	94.7	0.88	0.5
	C	33.3	24	1.38	
VENSES	E	50	47.01	1.08	0.16
	C	70	75.7	0.92	
WO	E	50	88	0.56	0.24
baseline	C	26.6	33	0.80	
ling-biased	E	96.6	98.5	0.98	0.03
baseline	C	40	39.4	1.01	

Table 5: Evaluation on entail. and contr. pairs

shallow approach implemented by the system has no strategies to correctly judge negative examples (similarly to the WO system), therefore should be mainly improved with this respect.

We also calculated the CI for every pair of the dataset, putting into relation each original pair with all the monothematic pairs derived from it. Figure 1 shows EDITS and VENSES's *CI* on each pair of our sample.[5] Even if the systems obtained similar performances in the challenge, the second system seems to behave in an opposite way with respect to EDITS, showing higher *CI* for negative cases than for the positive ones.

[5]The ideal case CI=1 corresponds to 0 on the logarithmic scale.

5 Conclusion and Future work

We have proposed a methodology for the evaluation of TE systems based on the analysis of the system behaviour on monothematic pairs with respect to the behaviour on corresponding original pairs. Through the definition of two indicators, a Correlation Index and a Deviation Index, we infer evaluation patterns which indicate strength and weaknesses of the system. As a pilot study, we have compared two systems that took part in RTE-5. We discovered that, although the two systems have similar accuracy on RTE-5 datasets, they show significant differences in their respective abilities to manage different linguistic phenomena and to properly combine them. We hope that the analysis provided by our methodology may bring interesting elements both to TE system developers and for deep discussion on the nature of TE itself.

As future work, we plan to refine the evaluation methodology introducing the possibility to assign different relevance to the phenomena.

6 Acknowledgements

Thanks to Professor Rodolfo Delmonte and to Sara Tonelli for running the VENSES system on our data sets.

References

Bentivogli, Luisa, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *Proceedings of the TAC 2009 Workshop on Textual Entailment*. 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 International Conference on Language Resources and Evaluation (LREC)* . Valletta, Malta. 19-21 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.

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

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

Gottlob, Frege. 1892. *Über Sinn und Bedeutung*. Zeitschrift für Philosophie und philosophische Kritik. 100.25-50.

Magnini, Bernardo, and Elena Cabrio. 2009. Combining Specialized Entailment Engines. *Proceedings of the 4th Language & Technology Conference (LTC '09)*. 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 Textual Entailment*. Gaithersburg, Maryland. 17 November.

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*. Volume 5883. ISBN 978-3-642-10290-5. Springer-Verlag Berlin Heidelberg, p. 314.

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.

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 the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*. Trento, Italy. 3-7 April.