Predicting Correlations Between Lexical Alignments and Semantic Inferences

Simone Magnolini
University of Brescia
FBK, Trento, Italy
magnolini@fbk.eu

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

Abstract

While there is a strong intuition that word alignments (e.g. synonymy, hyperonymy) play a relevant role in recognizing text-to-text semantic inferences (e.g. textual entailment, semantic similarity), this intuition is often not reflected in the system performances and there is a general need of a deeper comprehension of the role of lexical resources. This paper provides an empirical analysis of the dependencies between data-sets, lexical resources and algorithms that are commonly used in text-to-text inference tasks. We define a *resource impact index*, based on lexical alignments between pairs of texts, and show that such index is significantly correlated with the performance of different textual entailment algorithms. The result is an operational, algorithm-independent, procedure for predicting the performance of a class of available RTE algorithms.

1 Introduction

In the last decade text-to-text semantic inference has been a relevant topic in Computational Linguistics. Driven by the assumption that language understanding crucially depends on the ability to recognize semantic relations among portions of text, several text-to-text inference tasks have been proposed, including recognizing paraphrasing (Dolan and Brockett., 2005), recognizing textual entailment (RTE) (Dagan et al., 2005), and semantic similarity (Agirre et al., 2012). A common characteristic of such tasks is that the input are two portions of text, let's call them $Text1$ and $Text2$, and the output is a semantic relation between the two texts, possibly with a degree of confidence of the system. For instance, given the following text fragments:

Text1: *George Clooneys longest relationship ever might have been with a pig. The actor owned Max, a 300-pound pig.*
Text2: *Max is an animal.*

a system should be able to recognize that there is an ”entailment” relation among $Text1$ and $Text2$.

While the task is very complex, requiring in principle to consider syntax, semantics and also pragmatics, current systems adopt rather simplified techniques, based on available linguistic resources. For instance, many RTE systems (Dagan et al., 2012) would attempt to take advantage of the fact that, according to WordNet, the word *animal* in $Text2$ is a hypernym of the word *pig* in $Text1$. A relevant aspect in text-to-text tasks is that data-sets are usually composed of textual pairs for positive cases, where a certain relation (e.g. entailment) holds, and negative pairs, where the semantic relation does not hold. For instance, the following pair:

Text1: *John has a cat, named Felix, in his farm, it's a Maine Coon, it's the largest domesticated breed of cat.*
Text2: *Felix is the largest domesticated animal in John's farm.*

shows a case of ”non-entailment”. It is worth to notice that in both the examples, although the entailment judgment is different, still there is an high degree of lexical alignments between words in $Text1$ and $Text2$ (e.g. $Max \longrightarrow Max, pig \longrightarrow animal, cat \longrightarrow animal)$.

In the paper we systematically investigate the relations between the distribution of lexical associations in textual entailment data-sets and the system performance. As a result we define a ”resource impact index” for a certain lexical resource with respect to a certain data-set, which indicates the capacity of the resource to discrimi-

Proceedings of Recent Advances in Natural Language Processing, pages 388–397,
Hissar, Bulgaria, Sep 7–9 2015.

nate between positive and negative pairs. We show that the ”resource impact index” is homogeneous across several data-sets and tasks, and that it correlates with the performance of available entailment systems.

The paper is structured as follows. Section 2 provides the relevant background about the ongoing discussion on the use of lexical resources in textual entailment. Section 3 defines the Resource Impact Index that will be used in the experimental section. Section 4 reports on the experimental setting, including data-sets, resources and algorithms that we have been using. Section 5 discusses the results in term of the correlation between the Resource Index on a certain data-set and the accuracy obtained by two different algorithms using a single lexical relation at time. Section 6 shows how we can combine the Resource Index in case of multiple resources, while still maintaining the correlation with the algorithm performance. Finally, Section 7 highlights the potential impact of the paper within the current research on text-to-text semantic inferences.

2 Background on Lexical Resources and Text-to-Text Inferences

The role of lexical resources for recognizing text-to-text semantic relations (e.g. paraphrasing, textual entailment, textual similarity) has been under discussion for several years. This discussion is well reflected in the data reported by the RTE-5 ”ablation tests” initiative (Bentivogli et al., 2009), where the performance of an algorithm was measured removing one resource at a time.

Challenge	**T1/T2 Overlap (%)**		
	YES	**NO ENTAILMENT**	
		Unknown	**Contradiction**
RTE - 1	68.64	64.12	
RTE - 2	70.63	63.32	
RTE - 3	69.62	55.54	
RTE - 4	68.95	57.36	67.97
RTE - 5	77.14	62.28	78.93

Table 1: Comparison among the structure of different RTE data-sets (Bentivogli et al., 2009).

As an example, participants at the RTE evaluation reported that WordNet was useful (i.e. improved performance) 9 of the times, while 7 out of 16 it was not. In addition, Table 1, again extracted from (Bentivogli et al., 2009), suggests that the degree of word overlap among positive and negative pairs might be a key to understand the complexity of a text-to-text inference task, and, as a consequence, a key to interpret the system’s performance. Particularly, we can notice that the word overlap for the ”Yes” cases and the ”Contradiction” cases in the RTE-4 data-set is very similar, and even higher for the RTE-5 data-set. While this fact confirms the intuition that contradiction is generated when there is high overlap in meaning (de Marneffe, 2012), it also means that word overlap is not a discriminatory feature.

In this paper we claim that the two issues raised at RTE-5 (i.e. mixed evidence for the use of WordNet, and the fact that word overlap was not discriminative) are very much related, and, actually, are part of the same phenomenon. To support our claim, we build on top of previous work (Magnolini and Magnini, 2014), which we generalize considering: (i) lexical associations with different polarity (e.g. synonyms and antonyms); (ii) data-sets with different characteristics, (e.g. task, length of the pairs, languages); (iii) different algorithms for calculating textual entailment. We are interested to capture correlations between the use of lexical resources (both single resources and in combination) and the performance of inference algorithms. Particularly, the goal is to predict the behavior of an entailment algorithm given the characteristics of both the resource and the data-set.

There are several factors which, in principle, can affect our experiments, and that we have carefully considered.

Lexical Resources. First, the impact of a resource depends on the quality of the resource itself. Lexical resources, particularly those that are automatically acquired, might include noisy data, which can negatively affect performance. On the other hand, manually developed resources such as WordNet (Fellbaum, 1998) are particularly complex (i.e. a dozen of different relations, deep taxonomic structure, fine grained sense distinctions) and their use needs tuning. In order to face with these issues, we have selected manually constructed lexical resources, with a high degree of precision. In our experiments we have used lexical relations separately, in order to keep as much as possible under control their effect. Under this use, when we refer to a lexical resource we actually mean a resource that provides a specific lexical relation: for instance, a resource for lexical derivation, a resource for the hyperonymy relation, and so on. In addition, in the paper we consider both lexical resources that are supposed to provide similarity/compatibility alignments (e.g. synonyms) and resources/relations that are supposed to provide lexical oppositions (e.g. antonyms).

Inference Algorithms. Second, different algorithms may use different strategies to take advantage of resources. For instance, algorithms that calculate a distance or a similarity between $Text1$ and $Text2$ may assign different weights to a certain word association, on the basis on human intuitions (e.g. synonyms preserve entailment more than hypernyms). In our experiments we avoided as much as possible the use of settings not supported by empirical evidences and we use algorithms that are publicly available in order to maximize the replicability of the experiments.

Data-sets. Finally, data-sets representing different inference phenomena, may manifest different behaviors with respect to the impact of a certain resource, which can be specific for each inference type (e.g. entailment and semantic similarity). Although reaching a high level of generalization is limited by the existence of a limited number of data-sets, we have conducted experiments both on several textual entailment data-sets, also for different languages, and on a semantic similarity data-set.

3 Resource Impact Index

In this Section we define the general model through which we estimate the impact of a lexical resource. The idea behind the model is quite simple: the impact of a resource on a data-set should be correlated to the capacity of the resource to discriminate positive pairs from negative pairs in the data-set. We measure such capacity as the number of *lexical alignments* that the resource can establish on positive and negative pairs, and then we calculate the difference among them. We call this measure the *resource impact differential* - RID. The smaller the RID, the smaller the impact of the resource on that data-set. In the following we provide a more precise definition both of lexical alignments (Section 3.1) and of the model for calculating the resource impact differential (Section 3.2).

3.1 Defining Lexical Alignments

The idea that the entailment relation is related to the degree of lexical alignments between the words in a $(T1, T2)$ pair was introduced in (Dagan et al., 2012) as a useful generalization over the use of lexical resources in Recognizing Textual Entailment. In our work we adopt their definition of alignment, and we apply it to the RID calculation. More precisely, we say that two tokens in a $(T1, T2)$ pair are aligned when there is at least one semantic association relation, including equality, between the two tokens. For instance, synonyms and morphological derivations are different types of lexical alignments.

In addition, we extend the (Dagan et al., 2012) definition, allowing both positive and negative alignments. In fact, alignments inherit the polarity of the resource from which they are generated. We have a *Positive Alignment* when the semantic relation of the alignment is derived from a resource bringing positive associations (see Section 2), and we have a *Negative Alignment* when the source is negative (e.g. antonyms).

Finally, in the experiments reported in this paper we consider both word-to-word alignments and phrase alignments, where n-gram sequences are involved.

3.2 Defining the Impact Index

The Resource Impact Index is defined over a certain data-set D and a certain lexical resource LR.

Data-set (D). A data-set is a set of text pairs $D = \{(T1, T2)\}$, including both positive $(T1, T2)^p$ and negative $(T1, T2)^n$ pairs for a certain semantic relation (e.g. entailment, similarity). As reported in Section 2, this is a quite standard composition of benchmarks for text-to-text inferences.

Lexical Resource (LR). We define a Lexical Resource as any potential source of alignments among words. In most of the cases, rather than generic lexical resources (e.g. WordNet) we are interested in specific semantic relations provided by a resource. For instance, WordNet is a source for alignments based on synonyms. As discussed in Section 2, we consider both resources that are supposed to provide similarity-based alignments, which we call *positive lexical resources*, denoted with LR^+, and resources that are supposed to provide opposition-based alignments, which we call *negative lexical resources*, denoted with LR^-.

Resource Impact (RI). The impact of a resource LR on a data-set D is calculated as the number of lexical alignments returned by LR on all pairs, both positive and negative, normalized on the number of potential alignments for the data-set D. We use $|T1| * |T2|$ ($|T|$ is the number of tokens in text T) as potential number of potential alignments (Dagan et al., 2012, page 52), although there might be other options, such as $|T1| + |T2|$, and $max(|T1|, |T2|)$.

RI ranges from 0, when no alignment is found, to 1, when all potential alignments are returned by LR.

$$RI_{(LR,D)} = \frac{\sum_{i \in D} LexAl(T1_i, T2_i)}{\sum_{i \in D} |T1_i| * |T2_i|} \quad (1)$$

Resource Impact Differential (RID). The impact of a resource LR on a certain data-set D is given by the difference between the RI on positive pairs $(T1, T2) \in D^p$ and on negative pairs $(T1, T2) \in D^n$.

A RID for a positive lexical resource ranges from -1, when the RI is 0 for the positive pairs (i.e. when entailment holds) and 1 for negative entailed pairs, to 1, when the RI is 1 for entailed and 0 for not-entailed pairs.

$$RID_{(LR^+,D)} = RI_{(LR,D^p)} - RI_{(LR,D^n)} \quad (2)$$

For a resource with negative polarity (e.g. antonyms) the RID is expected to be the difference between the Resource Impact on negative and on positive pairs (equation 3).

$$RID_{(LR^-,D)} = RI_{(LR,D^n)} - RI_{(LR,D^p)} \quad (3)$$

The RID measure is not affected by the length of the pairs in the data-set, because it is normalized on the potential number of alignments for each pair. As far as the relation between RID and the impact of the lexical resource (i.e. the number of lexical alignments produced by the resource), being the RID a difference, we can consider the impact as an upper bound of the RID (see equation 4).

$$\left|RID_{(LR,D)}\right| \leq \frac{\sum_{i \in D} LexAl(T1_i, T2_i)}{\sum_{i \in D} |T1_i| * |T2_i|} \quad (4)$$

4 Experiments

In this section we apply the model described in Section 3 to different data-sets and resources, taking advantage of different sources of lexical and phrase alignments.

4.1 Data-sets

We use four different data-sets in order to experiment different characteristics of the corpora used for benchmarking text-to-text inferences.

RTE-3 eng. The RTE-3 data-set (Giampiccolo et al., 2007) for English has been used in the context of the Recognizing Textual Entailment shared tasks. It has been constructed mainly using application derived text fragments, and it is balanced between positive and negative pairs (about 1600 in total).

RTE-3 ita. The Italian RTE-3 data-set[1] is the translation of the English one. The goal is to monitor the behaviour of the RID while changing the language.

RTE-5 eng. The RTE-5 data-set (Bentivogli et al., 2009) is similar to RTE-3, although $T1$ pairs are usually much longer, which, in our terms, means that a higher number of alignments can be potentially generated by the same number of pairs.

SICK eng. Finally the SICK data-set (Sentences Involving Compositional Knowledge) (Marelli et al., 2014) has been recently used to highlight distributional properties. SICK is not balanced (1299 positive and 3201 negative pairs), and $T1$ and $T2$, differently from RTE pairs, have similar length.

4.2 Sources for Lexical Alignments

We carried out experiments using six different sources of lexical alignments, whose use is quite diffused in the practice of text-to-text inference systems, and with different expected behavior, as far as the polarity of the lexical resource is concerned.

Lemmas. The first source consists of a simple match among the lemmas in $T1$ and $T2$: if two lemmas are equal (case insensitive), then we count it as an alignment between $T1$ and $T2$. The expected polarity of alignments based on lemmas is positive, as we assume that they increase the similarity between $T1$ and $T2$.

[1] http://www.excitement-project.eu/index.php/results/178-public-resources

Synonyms. The second source considers alignments due to the synonymy relation (e.g. *home* and *habitation*). The sources are WordNet (Fellbaum, 1998), version 3.0 for English, and MultiWordNet (Pianta et al., 2002) for Italian. If two lemmas are found in the same synset, then we count it as an alignment.The expected polarity of alignments based on synonyms is positive.

Hypernyms. The third source considers the hyperonymy relation (e.g. *dog* and *mammal*): as for synonymy we use WordNet and MultiWordNet, counting as an alignment all the cases where two lemmas are in the hypernym hierarchy, at any distance. The expected polarity of alignments based on hypernyms is positive.

Morphological Derivations. The fourth source of alignment are morphological derivations (e.g. *invention* and *invent*). As for English, derivations are covered again by WordNet, while for Italian we used MorphoDerivIT, a resource developed within the EXCITEMENT project[2], which has the same structure of CATVAR (Habash and Dorr, 2003) for English. The expected polarity of alignments based on morphological derivations is positive.

Antonyms. The fifth source of alignment are antonyms (e.g. *man* and *woman*). Antonyms are provided by WordNet for English and by MultiWordNet for Italian. The expected polarity of alignments based on antonyms is negative, as we assume that they increase the opposition between $T1$ and $T2$.

Paraphrase Tables. The sixth source of alignment are paraphrase tables (e.g. *can be modified* and *may be revised*). We built paraphrase tables from the Meteor translation tables (Denkowski and Lavie, 2014). The idea is that if an n-gram n_s in the source language s is translated into n-gram n_t in the target language t, and if n_t has multiple translations back into s, then all these translations are potential paraphrases of each other. The probability of translation from one language to another can be used to compute the probability that two n-grams in language s are paraphrases of each other. To compute this probability we use all shared translations into the target language t of the two n-grams (both in source language s). There are two main reasons to consider paraphrase tables: (i) they cover alignments that are only partially covered by the other sources that we considered; (ii) most of the phrases are n-grams, which allows us to test the RID behavior on sequences longer than single tokens. The expected polarity of alignments based on paraphrase tables is positive.

0-Knowledge. Finally, in order to investigate the behavior of the RID in absence of any lexical alignment, we include a 0-Knowledge experimental baseline, where the system does not have access to any source of lexical alignment. As no alignment is produced (including token match), the RID of the 0-Knowledge baseline is always 0.

4.3 Algorithms

In order to verify our hypothesis that the RID index is correlated with the capacity of a system to correctly recognize textual entailment, we run experiments using two different RTE algorithms, i.e. EDITS and P1EDA, which take advantage of lexical resources in different ways. The two algorithms are both supervised, in the sense that they use training data to build a model. As the goal of our experiments is to monitor the behavior of the RID index in different settings, rather than to assess the performance of the two algorithms, we decided to simplify as much as possible the experimental setting, and we calculated accuracy and F1 for the two algorithms using the training section of the data-sets[3].

EDITS (Negri et al., 2009), is a distance-based RTE algorithm based on calculating the Edit Distance between $T1$ and $T2$, defined as the minimum-weight sequence of edit operations (i.e. deletion, insertion and substitution) that transforms $T1$ into $T2$. The intuition is that the less the cost of transforming $T1$ into $T2$, the more likely the entailment relation between the two texts. The final decision is taken on the basis of a threshold, empirically estimated over training data. For all the experiments, the cost of edit operations is set as follows: 0 for substitution if two words are aligned; 1 for substitution if two words are not aligned; 1 for insertion; 0 for deletion. The algorithm is normalized on the number of words of $T1$ and $T2$, after stop

[2] http://www.excitement-project.eu/index.php/results/178-public-resources

[3] We will investigate the behavior of the RID between test and training data-sets in future work.

words are removed. As for linguistic processing, the Edit Distance algorithm needs tokenization, lemmatization and Part-of-Speech tagging (in order to access resources). We used TreeTagger (Schmid, 1995) for English and TextPro (Pianta et al., 2008) for Italian. In addition we removed stop words, including some very common verbs.

P1EDA (Noh et al., 2015) is an alignment-based RTE algorithm, developed and fully documented in the software website[4], based on alignments between $T1$ and $T2$. The intuition is that the more the portions of $T2$ are aligned with portions of $T1$, the higher the probability of the entailment relation. First the algorithm extracts all possible alignments between portions in $T1$ and $T2$, then it extracts a number of features from the alignments, which are finally given as input to a multinomial logistic regression classifier trained on annotated data. The features implemented in the P1EDA version used for our experiments are the following: (i) the ratio of words in $T2$ aligned with $T1$; (ii) the ratio of content words in $T2$ aligned with $T1$ and, (iii) the ratio of verbs in $T2$ aligned with $T1$. As for linguistic processing, P1EDA needs tokenization, lemmatization and Part-of-Speech tagging. As in the case of EDITS we used TreeTagger (Schmid, 1995) for English and TextPro (Pianta et al., 2008) for Italian.

All the experiments reported in the paper have been conducted using the Excitement Open Platform (EOP), (Padó et al., 2014) (Magnini et al., 2014), a generic architecture and a comprehensive implementation for textual inference in multiple languages. The platform includes state-of-art algorithms, a large number of knowledge resources and facilities for experimenting and testing innovative approaches. The architecture is based on the concept of modularization with pluggable and replaceable components to enable extensions and customizations, this way helping to control that experiments are conducted in the proper way, with easily observable intermediate steps. The EOP platform includes both the algorithms and the lexical resources used in our experiments, and it is distributed as an open source software.[5]

5 Results

Table 2 and Table 3 report the results of the experiments on the four data-sets and the seven sources of alignment (including the 0-Knowledge baseline) described in Section 4[6]. For each resource we show the RID of the resource (given the very low values, RIDs are shown multiplied by a 10^4 factor), and the accuracy achieved both by the EDITS and the P1EDA algorithms. The last row of the tables shows the Pearson correlation between the RID and the accuracy of the algorithms for each data-set, calculated as the mean of the correlations obtained for each resource on that data-set.

A first observation is that all RID values are very close to 0, indicating a low expected impact of the resources. Even the highest RID (i.e. 523.342 for lemmas on SICK), corresponds to a 5% of the potential impact of the resource. Negative RID values for positive resources, mean that the resource, somehow contrary to the expectation, produces more alignments for negative pairs than for positive (this is the case, for instance, of synonyms on the English RTE-3). On the same line, negative RID values for negative resources mean that a resource with negative polarity produces more alignments for positive pairs than for negative (this case does not appear in the results).

Alignment on lemmas is by far the resource with the best impact, while alignments produced by paraphrases produce very negative RID.

Finally, results fully confirm the initial hypothesis that the RID is correlated with the system performance; i.e. the accuracy for balanced data-sets and the F1 for the unbalanced one. The Pearson correlation shows that R is close to 1 for all the RTE data-sets (the slightly lower value on SICK reveals the different characteristics of the data-set), indicating that the RID is a very good predictor of the system performance, at least for the class of inference algorithms represented by EDITS and P1EDA. The low values for RID are also reflected in absolute low performance, showing again that when the system uses a low impact resource the accuracy is close to the baseline (i.e. the 0-Knowledge configuration).

Although improving the performance of RTE systems is not the direct goal of our experiments, it is worth noting that P1EDA outperformed EDITS,

[4] https://github.com/hltfbk/EOP-1.2.3/wiki/AlignmentEDAP1

[5] http://hltfbk.github.io/Excitement-Open-Platform/

[6] The EDITS implementation available in the EOP platform does not allow n-gram alignments, so we could not run paraphrases with EDITS.

EDITS	RTE-3 eng		RTE-3 ita		RTE-5 eng		SICK eng	
	RID	Accuracy	RID	Accuracy	RID	Accuracy	RID	F1
0-Knowledge	0	0.542	0	0.543	0	0.536	0	0.004
Lemmas	97.215	0.635	84.594	0.641	43.221	0.62	523.342	0.347
Synonyms	-4.876	0.536	5.343	0.537	10.138	0.561	12.386	0.093
Hypernyms	-5.333	0.532	-1.791	0.543	12.921	0.555	48.665	0.221
Derivations	-1.747	0.571	-0.024	0.536	5.722	0.553	-6.436	0
Antonyms (*)	1.076	0.542	0	0.543	1.013	0.54	28.479	0
R Correlation	0.943		0.990		0.988		0.862	

Table 2: Experimental results on different data-sets with different resources using EDITS. (*) Antonyms have negative polarity.

P1EDA	RTE-3 eng		RTE-3 ita		RTE-5 eng		SICK eng	
	RID	Accuracy	RID	Accuracy	RID	Accuracy	RID	F1
0-Knowledge	0	0.527	0	0.517	0	0.506	0	0
Lemmas	97.215	0.682	84.594	0.706	43.221	0.601	523.342	0.485
Synonyms	-4.876	0.533	5.343	0.516	10.138	0.521	12.386	0
Hypernyms	-5.333	0.527	-1.791	0.512	12.921	0.543	48.665	0.038
Derivations	-1.747	0.553	-0.024	0.512	5.722	0.528	-6.436	0.018
Antonyms (*)	1.076	0.532	0	0.517	1.013	0.52	28.479	0
Paraphrases	-11.668	0.52	18.049	0.5075	33.803	0.563	-67.148	0.015
R Correlation	0.987		0.967		0.959		0.983	

Table 3: Experimental results on different data-sets with different resources using P1EDA. (*) Antonyms have negative polarity.

	RID$_C$	Accuracy (P1EDA)	R Correlation
0-knowledge	0	0.527	
Lemmas+Synonyms	92.338	0.683	
Synonyms+Hypernyms	-10.209	0.526	
Hypernyms+Antonyms	-6.409	0.528	
			0.996
ALL resources	84.181	0.687	
			0.995
Paraphrases+Synonyms	-16.296	0.523	
			0.993

Table 4: Results on combining multiple resources using P1EDA.

and it achieved results (i.e. 0.68 on English RTE-3, 0.70 on Italian RTE-3, 0.60 on RTE-5) which can be considered at the state-of-art for publicly available systems.

6 Combining RIDs of Multiple Sources

While the previous sections have confirmed our hypothesis that the RID index is correlated with the performance of RTE algorithms using single resources, the aim of this Section is to show that the RID obtained from a combination of resources is still correlated with the algorithm performance.

We define the RID of multiple resources, called Combined Resource Index Differential (RID_C) as the sum of the RIDs of the single resources. For instance, in Table 4, the combined RID_C of Lemmas+Synonyms (i.e. 92.338) is obtained summing the RID for Lemmas (97.215, see Table 3) with the RID for Synonyms (i.e. -4.876). Intuitively, the sum of two RIDs for the resources $LR1$ and $LR2$ corresponds to the RID

of a single resource composed by $LR1$ and $LR2$, under the assumption that they are disjoint, i.e. that the set of alignments that $LR1$ and $LR2$ produce is disjoint. In order to take into consideration the combination of non-disjoint resources, the RID of the intersection has to be subtracted, as shown in equation 5 (combining positive resources) and equation 6 (combining a positive and a negative resource).

$$\begin{aligned} RID_{C(LR_1^+, LR_2^+, D)} = RID_{(LR_1^+, D)} + \\ RID_{(LR_2^+, D)} - RID_{(LR_1^+ \cap LR_2^+, D)} \end{aligned} \tag{5}$$

$$\begin{aligned} RID_{C(LR_1^+, LR_2^-, D)} = RID_{(LR_1^+, D)} - \\ RID_{(LR_2^-, D)} - RID_{(LR_1^+ \cap LR_2^-, D)} \end{aligned} \tag{6}$$

We conducted a number of RID combination experiments, reported in Table 4. First, we used four disjoint resources, whose RIDs show different characteristics on the RTE-3 dataset. As reported in Table 3, lemmas have a high and positive RID; synomyms and hypernyms are both resources with positive polarity, and both have a slightly negative RID; antonyms is a resource with negative polarity and slightly positive RID. For each pairwise combination, we run P1EDA for calculating entailment judgments, and then we computed the correlation between the accuracy of the algorithm and the RID of the combination, calculated summing the RIDs.

Then, we experimented a combination of the five resources (including the 0-Knowledge baseline). The result (”All resources” line in Table 4), again shows very high correlation with the accuracy of the system. We think that the minor decrease in the correlation (i.e. from 0.996 to 0.995) is due to few cases of overlap among the resources, particularly some synonyms are also hypernyms, which we did not filter out.

Finally, we run a combination experiment using paraphrases and synonyms, two resources that show a relatively high level of overlap in RTE-3. Here the goal is to test that subtracting the RID of the intersection of the two resources results in a better correlation. Accordingly, we have calculated both the simple RID (i.e. without subtracting the RID of the intersection) and the combined RID_C. We note that the alignments in the intersection are almost equally distributed between positive and negative pairs, resulting in very close RIDs, namely -16.544 for the simple RID, and -16.296 for the combined one.

7 Final Discussion and Conclusion

According to the initial working hypothesis, we have shown that the RID index is highly correlated with the accuracy of RTE systems, a result that allows to use the RID as a reliable indicator of the impact both of a single resource and of a combination of them. We now have both an empirical explanation of the impact of a lexical resource over a certain inference task, and an operational, algorithm-independent procedure for predicting the performance of a class of available RTE algorithms.

We now discuss what we can learn from the achievements reported in the paper, and how we can take advantage of our findings in order to design more effective text-to-text inference systems.

A first finding is that RIDs of popular lexical relations among words are quite close to 0, which indicates that their distribution is not useful to discriminate positive and negative pairs in current text-to-text data-sets. As a second finding, the Resource Impact RI (equation 1), which tells us how much a resource is used for a certain data-set, is very dis-homogeneous. To give an idea, the following are the RIs of our resources on the English RTE-3 data-set: lemmas 682.266, synonyms 72.709, hypernyms 157.055, morphological derivations 62.757, antonyms 3.885, paraphrases 316.717. Finally, although we do not have quantitative data supporting our intuition, we are convinced that the coverage of our resources (i.e. the alignments produced by a resources with respect to the alignments it should produce) is pretty good, indicating that there is no much room for improving the resources themselves.

Given the above three elements, i.e. low RID of resources (even in combination), not homogeneous impact of different semantic relations, and good coverage over the data-sets, we think that future improvements in text-to-text inference should consider more discriminative features, i.e. resources with higher absolute value of RID (e.g. a wider range of lexical opposition phenomena). In addition, our findings support the intuition that lexical phenomena do not exhaust the complexity of textual entailment and that local compositional aspects of meaning (e.g. verb argument structure, scope of negation), need to be exploited.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 385–393, Montréal, Canada.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognising textual entailment challenge. In *Proceedings of the TAC Workshop on Textual Entailment*, Gaithersburg, MD.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 177–190, Southampton, UK.

Ido Dagan, Dan Roth, and Fabio Massimo Zanzotto. 2012. *Recognizing Textual Entailment: Models and Applications*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Marie-Catherine de Marneffe. 2012. *What's that supposed to mean?* Ph.D. thesis, Stanford Univeristy.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005), Asia Federation of Natural Language Processing*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for english. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 17–23. Association for Computational Linguistics.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics, Demo papers*.

Simone Magnolini and Bernardo Magnini. 2014. Estimating lexical resources impact in text-to-text inference tasks. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Italy.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Matteo Negri, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. 2009. Towards extensible textual entailment engines: the EDITS package. In *Proceeding of the Conference of the Italian Association for Artificial Intelligence*, pages 314–323, Reggio Emilia, Italy.

Tae-Gil Noh, Sebastian Padó, Vered Shwartz, Ido Dagan, Vivi Nastase, Kathrin Eichler, Lili Kotlerman, and Meni Adler. 2015. Multi-level alignments as an extensible representation basis for textual entailment algorithms. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*. `doi:10.1017/S1351324913000351`.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Intl Conference on Global WordNet*.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Helmut Schmid. 1995. Treetagger - a language independent part-of-speech tagger. *Insti-*

tut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43:28.