10.12871/CLICIT2014147

Estimating Lexical Resources Impact in Text-to-Text Inference Tasks

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

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

Abstract

English. This paper provides an empirical analysis of both the datasets and the lexical resources that are commonly used in text-to-text inference tasks (e.g. textual entailment, semantic similarity). According to the analysis, we define an index for the impact of a lexical resource, and we show that such index significantly correlates with the performance of a textual entailment system.

Italiano. *Questo articolo fornisce un'analisi empirica dei datasets e delle risorse lessicali comunemente usate per compiti di inferenza testo-a-testo (es., implicazione testuale, similaritá semantica). Come risultato definiamo un indice che misura l'impatto di una risorsa lessicale, e mostriamo che questo indice correla significativamente con le prestazioni di un sistema di implicazione testuale.*

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 datasets are usually composed of textual pairs for positive cases, where a certain relation occurs, and negative pairs, where a semantic relation doesn't appear. 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".

In the paper we systematically investigate the relations between the distribution of lexical associations in textual entailment datasets and the system performance. As a result we define a "resource impact index" for a certain lexical resource with respect to a certain dataset, which indicates the capacity of the resource to discriminate between positive and negative pairs. We show that the "resource impact index" is homogeneous across several datasets and tasks, and that it correlates with the performance of the algorithm we chose in our experiments.

2 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 since several years. This discussion is well reflected in the data reported by the RTE-5 ”ablation tests” (Bentivogli et al., 2009), where the performance of a certain algorithm was measured removing one resource at 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-set (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 of the time it wasn't useful. As an initial explanation for such controversial behavior, 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. In this paper we extend this intuition, considering: (i) lexical associations (e.g. synonyms) other than word overlap, and (ii) datasets with different characteristics.

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

Resource. 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 negatively affect performance. In addition, resources such as WordNet (Fellbaum, 1998) are particularly complex (i.e. dozen of different relations, deep taxonomic structure, fine grained sense distinctions) and their use needs tuning. We have selected lexical resources manually constructed, with a high degree of precision, and in the experiments we have used lexical relations separately, in order to keep under control their effect.

Inference Algorithm. 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 grounded on empirical evidences.

Dataset. Finally, datasets representing different inference phenomena, may manifest different behaviors with respect to the impact of a certain resource, specific for each inference type (e.g. entailment and semantic similarity). Although reaching a high level of generalization is limited by the existence itself of datasets, we have conducted experiments both on textual entailment and semantic similarity.

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 dataset should be correlated to the capacity of the resource to discriminate positive pairs from negative pairs in the dataset. We measure this capacity in term of 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 dataset. In the following we provide a more precise definition of the model.

Dataset (D). A dataset is a set of text pairs $D = \{(T1, T2)\}$, with positive $(T1, T2)^p$ and negative $(T1, T2)^n$ pairs for a certain semantic relation (e.g. entailment, similarity).

Lexical Alignment (LexAl). We say that two tokens in a $(T1, T2)$ pair are aligned when there's some semantic association relation, including equality, between the two tokens. For instance, synonyms and morphological derivations are different types of lexical alignments.

Lexical Resource (LR). A Lexical Resource is a potential source of alignment among words. For instance, WordNet is a source for synonyms [1].

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, normalized on the number of potential alignments for the data-set D. We use $|T1| * |T2|$ as potential alignments (Dagan et al., 2012, page 52), although there might be other options: $|T1| + |T2|$, $max(|T1|, |T2|)$, etc. RI ranges from 0, when no alignment is found, to 1, when all potential alignments are returned by LR.

$$RI_{(LR,D)} = \#LexAl/|T1| * |T2| \qquad (1)$$

Resource Impact Differential (RID). The impact of a resource LR on a certain dataset D is given by the difference between the RI on positive pairs $(T1, T2)^p$ and on negative pairs $(T1, T2)^n$. A RID ranges from -1, when the RI is 0 for the entailed pairs and 1 for not entailed pairs, to 1, when the RI is 1 for entailed and 0 for not entailed pairs.

$$RID_{(LR,D)} = RI(T1, T2)^p - RI(T1, T2)^n \quad (2)$$

The RID measure isn't affected by the size of the dataset, because it's normalized on the maximum number of alignments. Finally, the coverage of the resource (i.e. the number of lexical alignments) is an upper of the bound of the RID (see 3), being the RID a difference.

$$\left|RID_{(LR,D)}\right| \leq \frac{\#LexAl}{|T1| \cdot |T2|} \qquad (3)$$

4 Experiments

In this section we apply the model described in Section 3 to different datasets and resources, showing that the RID is highly correlated to the accuracy of a text-to-text inference algorithm.

Datasets. We use four different datasets in order to experiment different characteristics of text-to-text inferences. The RTE-3 dataset (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's balanced between positive and negative pairs (about 1600 in total). The Italian RTE-3 dataset is the translation of the English one. The RTE-5 dataset is similar to RTE-3, although Text-1 in 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. Finally the SICK dataset (Sentences Involving Compositional Knowldedge) (Marelli et al., 2014) has been recently used to highlight distributional properties, it isn't balanced (1299 positive and 3201 negative pairs), and $T1$ and $T2$, differently from RTE pairs, have similar length.

Sources for lexical alignments. We carried on experiments using four different sources of lexical alignments, whose use is quite diffused in the practice of text-to-test inference systems. 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 second resource considers alignments due to the synonymy relation (e.g. *home* and *habitation*). The source is WordNet (Fellbaum, 1998), version 3.0 for English, and MultiWordNet (Pianta et al., 2002) for Italian. The third resource considers the hypernym relation (e.g. *dog* and *mammal*): as for synonymy we use WordNet. The last source of alignment are morphological derivations (e.g. *invention* and *invent*). For English derivations are covered again by WordNet, while for Italian we used MorphoDerivIT, a resource developed at FBK which has the same structure of CATVAR (Habash and Dorr, 2003) for English. 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.

Algorithm. 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 all the experiments using EDITS (Negri et al., 2009) RTE based on calculating the Edit Distance between $T1$ and $T2$ in a pair. The algorithm calculate the minimum-weight series of edit operations (deletion, insertion and substitution) that transforms $T1$ into $T2$.The algorithm has an optimizer that decides the best cost for every edit op-

[1] In the paper we consider lexical resources that are supposed to provide similarity/compatibility alignments (e.g. synonyms). However, there might be resources (e.g. antonyms in WordNet) that are supposed to provide dissimilarity/opposition alignments. We'll investigate negative alignments in future work.

	RTE-3 eng		RTE-3 ita		RTE-5 eng		SICK eng	
	RID	Accuracy	RID	Accuracy	RID	Accuracy	RID	F1
0-Knowledge	0	0.537	0	0.543	0	0.533	0	0.005
Lemmas	87.164	0.617	84.594	0.641	36.169	0.6	523.342	0.347
Synonyms	-6.432	0.533	5.343	0.537	1.383	0.546	12.386	0.093
Hypernyms	-0.017	0.545	-1.790	0.543	7.969	0.556	48.665	0.221
Derivations	0.154	0.543	-0.024	0.536	2.830	0.545	-6.436	0
R correlation	0.996		0.991		0.985		0.851	

Table 2: Experimental results obtained on different datasets with different resources.

erations. The algorithm is normalized on the number of words of $T1$ and $T2$ after stop 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 (Emanuele Pianta and Zanoli, 2008) for Italian. In addition we removed stop words, including some of the very common verbs. Finally, all the experiments have been conducted using the EXCITEMENT Open Platform (EOP) (Padó et al., 2014) (Magnini et al., 2014), a rich and modular open source software environment for textual inferences [2].

5 Results and Discussion

Table 2 reports the results of the experiments on the four datasets and the five sources of alignment (including the 0-Knowledge baseline) described in Section 4. For each resource we show the RID of the resource (given the very low values, the RID is shown multiplied by a 10^4 factor), and the accuracy achieved by the EDITS algorithm. The last row of the Table shows the Pearson correlation between the RID and the accuracy of the algorithm for each dataset, calculated as the mean of the correlations obtained for each resource on that dataset.

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 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). Alignment on lemmas is by far the resource with the best impact.

Finally, results fully confirm the initial hypothesis that the RID is correlated with the system performance; i.e. the accuracy for balanced datasets and the F1 for the unbalanced one. The Pearson correlation shows that R is close to 1 for all the RTE datasets (the slightly lower value on SICK reveals the different characteristics of the dataset), indicating that the RID is a very good predictor of the system performance, at least for the class of inference algorithms represented by EDITS. 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).

6 Conclusion

We have proposed a method for estimating the impact of a lexical resource on the performance of a text-to-text semantic inference system. The starting point has been the definition of the RID index, which captures the intuition that in current datasets useful resources need to discriminate between positive and negative pairs. We have then shown that the RID index is highly correlated with the accuracy of the system for balanced datasets and with the F1 for the unbalanced one, a result that allows to use the RID as a reliable indicator of the impact of a resource.

As for future work, we intend to further generalize our current findings applying the same methodology to different text-to-text inference algorithms, starting from those already available in the EXCITEMENT Open Platform. We also want to conduct experiment on operation, like summing, with this index to describe to combined effect of different resources.

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

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.

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*.

Christian Girardi Emanuele Pianta 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/.

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*.

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.

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*.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.